Householder Symposium XVIII on Numerical Linear Algebra

June 12-17, 2011
Granlibakken Conference Center & Lodge
Tahoe City, California
Contents

Householder Symposium XVIII on Numerical Linear Algebra 1

Householder Committee 4

Local Organizer 4

Householder Prize Committee 4

Acknowledgments 5

Abstracts 6

Awad H. Al-Mohy  
An Improved Algorithm for the Matrix Logarithm 7

David Amsallem  
Interpolation on Matrix Manifolds of Reduced-Order Models and Application to On-Line Aeroelastic Predictions 9

Athanasios C. Antoulas  
Model Reduction of Parameter-Dependent Systems 10

Mario Arioli  
Generalized Golub-Kahan Bidiagonalization and Stopping Criteria 12

Haim Avron  
Randomized Algorithms in Numerical Linear Algebra: From Theory to Practice 14

Zhaojun Bai  
Progress in Linear and Nonlinear Eigensolvers 16

Grey Ballard  
Avoiding Communication using Successive Band Reduction 17

Ann-Kristin Baum  
Positivity Preserving Discretizations for Differential-Algebraic Systems 19

Christopher Beattie  
Model Reduction of Hamiltonian Systems in Variational Data Assimilation 21

Peter Benner  
Rational Krylov Subspaces for Nonlinear Model Reduction 23

Michele Benzi  
Computation of Matrix Functions Arising in Graph Analysis 25

Paolo Bientinesi  
A Modular and Systematic Approach to Stability Analysis 26

David Bindel  
A Matrix Factorization for Computer Network Tomography 27

Daniel Boley  
Commuting Times for a Directed Graph using an Asymmetric Laplacian 28
Shreemayee Bora
Moving Eigenvalues with Structured Perturbations, Crawford Number and ε-pseudospectra .................................................. 30

Rüdiger Borsdorf
Computing a Nearest Correlation Matrix with Factor Structure ................... 32

Mike A. Botchev
Residual, Restarting and Richardson Iteration for the Matrix Exponential ........ 34

Karen Braman
Theoretical Foundations for Computation of Eigendecompositions of Third-Order Tensors ......................................................... 35

Alfredo Buttari
Fine Grained Sparse QR Factorization for Multicore Systems .................... 37

Russell Carden
Ritz Values of Nonsymmetric Matrices and Convergence of the Restarted Arnoldi Method .......................................................... 39

Younès Chahlaoui
Gramian Based Model Reduction of Switched Dynamical Systems .............. 41

Shivkumar Chandrasekaran

Xiao-Wen Chang
LLL Reduction and Integer Least Squares Problems ................................ 44

Saifon Chaturantabut
Error Analysis for Nonlinear Model Reduction using POD-DEIM Technique 46

Julianne Chung
Windowed Spectral Regularization of Inverse Problems ............................ 47

Achiya Dax
The Eckart-Young Theorem and Ky Fan’s Maximum Principle: Two Sides of the Same Coin ....................................................... 49

James Demmel
Avoiding Communication in Numerical Linear Algebra .............................. 50

Eric de Sturler
Convergence Bounds for Approximate Invariant Subspace Recycling for Sequences of Linear Systems ................................................ 51

Fernando de Terán
Congruence Orbits of Matrices and Palindromic Pencils, and the Solution of the Equation XA + AXT = 0 ............................................. 53

Inderjit Dhillon
Social Network Analysis: Fast and Memory-Efficient Low-Rank Approximation of Massive Graphs .............................................. 55

Marco Donatelli
Square Smoothing Operators Imposing Accurate Boundary Conditions .......... 57
Froilán M. Dopico
Structured Perturbation Theory of Diagonally Dominant Matrices and Numerical Applications ........................................... 59

Zlatko Drmač
Numerical Implementation of the Iterative Rational Krylov Algorithm for Optimal $H_2$ Model Reduction .................................. 60

Vladimir Druskin
Optimal Krylov Subspaces for Large-Scale Dynamical Systems .......... 61

Iain Duff
Preconditioners based on Strong Components ................................ 63

Jurjen Duintjer Tebbens
On Arbitrary Convergence Behavior of the Arnoldi Method .............. 65

Ioana Dumitriu
Randomized Algorithms for Communication-Optimal SVD and EIG ........ 67

Alan Edelman
Needle-like Triangles, Matrices and Lewis Carroll .......................... 69

Lars Eldén
Computing Low-Rank Approximations of Sparse Tensors using Krylov Methods, and Applications in Information Sciences ........ 70

Howard C. Elman
Linear Algebra Problems Arising from Parameter-Dependent Partial Differential Equations ................................................. 71

Mark Embree
The Stability of GMRES Convergence with Applications to Inexact Preconditioning . 72

Yogi Ahmad Erlangga
Multilevel Krylov Method for the Biharmonic Equation ..................... 74

Oliver G. Ernst
Krylov Subspace Methods for Geoelectrical Exploration Problems ........ 76

Heike Fassbender
Numerical Linear Algebra in POD for Model Order Reduction of Steady Aerodynamic Applications ........................................ 77

David Chin-lung Fong
LSMR: An Iterative Algorithm for Least-squares Problems .................. 78

Leslie V. Foster
Reliable Calculation of Numerical Rank, Null Space Bases, Basic Solutions and Pseudoinverse Solutions using SuiteSparseQR ............... 79

Melina A. Freitag
New Algorithms for Calculating the Distance to Instability and the Distance to a Nearby Defective Matrix .............................. 80

Shmuel Friedland
Theoretical and Numerical Results and Problems in Tensors ............... 82
<table>
<thead>
<tr>
<th>Author</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andreas Frommer</td>
<td>Computational Proofs of the Stability of Lyapunov Equations</td>
<td>84</td>
</tr>
<tr>
<td>David F. Gleich</td>
<td>Skew-symmetric Matrix Completion for Ranking</td>
<td>86</td>
</tr>
<tr>
<td>Anne Greenbaum</td>
<td>Bounds on Norms of Functions of Matrices Using the Field of Values</td>
<td>87</td>
</tr>
<tr>
<td>Chen Greif</td>
<td>Towards Block Structure-Preserving Saddle-Point Solvers</td>
<td>88</td>
</tr>
<tr>
<td>Laura Grigori</td>
<td>CALU: A Communication Optimal LU Factorization Algorithm</td>
<td>89</td>
</tr>
<tr>
<td>Luka Grubišić</td>
<td>Basic Iterative Algorithms of Numerical Linear Algebra as Building Blocks for Hybrid Adaptive Finite Element Methods</td>
<td>91</td>
</tr>
<tr>
<td>Ming Gu</td>
<td>Reduced Rank Regression via Convex Optimization</td>
<td>93</td>
</tr>
<tr>
<td>Stefan Güttel</td>
<td>Towards Black-Box Rational Krylov Methods for $f(A)b$ : Automated Parameter Selection for Markov Functions and Error Estimation</td>
<td>95</td>
</tr>
<tr>
<td>Serkan Gugercin</td>
<td>Optimal $H_2$ Points in action: From Model Reduction to Lebesgue Constant</td>
<td>97</td>
</tr>
<tr>
<td>Chun-Hua Guo</td>
<td>On a Nonlinear Matrix Equation Arising in Nano Research</td>
<td>99</td>
</tr>
<tr>
<td>Mert Gurbuzbalaban</td>
<td>A Fast Algorithm for Approximating the Distance to Instability</td>
<td>100</td>
</tr>
<tr>
<td>Martin H. Gutknecht</td>
<td>Spectral Deflation in Krylov Solvers</td>
<td>102</td>
</tr>
<tr>
<td>Christopher J. Hillar</td>
<td>Spectral Relaxations of Hard Combinatorial Problems</td>
<td>104</td>
</tr>
<tr>
<td>Iveta Hnětynková</td>
<td>The Golub-Kahan Iterative Bidiagonalization in Regularization of Ill-posed Problems and Estimation of the Noise in the Data</td>
<td>106</td>
</tr>
<tr>
<td>Mark Hoemmen</td>
<td>Communication-avoiding Krylov Subspace Methods</td>
<td>108</td>
</tr>
<tr>
<td>Thomas Huckle</td>
<td>Computations in Quantum Tensor Networks</td>
<td>110</td>
</tr>
<tr>
<td>Bruno Iannazzo</td>
<td>A Subspace Shift Technique for Nonsymmetric Algebraic Riccati Equations</td>
<td>111</td>
</tr>
<tr>
<td>Ilse Ipsen</td>
<td>Numerical Accuracy and Sensitivity of Monte Carlo Matrix Multiplication</td>
<td>113</td>
</tr>
<tr>
<td>Author</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Elias Jarlebring</td>
<td>Solving Nonlinear Eigenvalue Problems with a Linear Arnoldi Method in a Function Setting and a Characterization of Associated Invariant Pairs</td>
<td>114</td>
</tr>
<tr>
<td>Pavel Jiránek</td>
<td>A Posteriori Error Estimates Including Algebraic Error and Stopping Criteria for Iterative Solvers</td>
<td>116</td>
</tr>
<tr>
<td>Stefan Johansson</td>
<td>Stratification of Full Normal Rank Polynomial Matrices</td>
<td>118</td>
</tr>
<tr>
<td>Bo Kågström</td>
<td>Efficient and Reliable Algorithms for Challenging Matrix Computations Targeting Multicore Architectures and Massive Parallelism</td>
<td>120</td>
</tr>
<tr>
<td>William Kahan</td>
<td>Refining the General Symmetric Eigenproblem</td>
<td>122</td>
</tr>
<tr>
<td>Lars Karlsson</td>
<td>Parallel Two-Stage Reduction to Hessenberg Form using Shared Memory</td>
<td>123</td>
</tr>
<tr>
<td>Linda Kaufman</td>
<td>Modifications of an Algorithm for Factoring Symmetric Banded Indefinite Matrices</td>
<td>125</td>
</tr>
<tr>
<td>Misha E. Kilmer</td>
<td>Approximations of Third Order Tensors as Sums of (Non-negative) Low-rank Product-Cyclic Tensors</td>
<td>126</td>
</tr>
<tr>
<td>Carl Christian Kjelgaard Mikkelsen</td>
<td>Incomplete Cyclic Reduction for Narrow Banded and Diagonally Dominant Linear Systems</td>
<td>128</td>
</tr>
<tr>
<td>Tamara G. Kolda</td>
<td>On the Best Symmetric Rank-k Approximation of a Symmetric Tensor</td>
<td>130</td>
</tr>
<tr>
<td>Daniel Kressner</td>
<td>Bivariate Matrix Functions</td>
<td>131</td>
</tr>
<tr>
<td>Julien Langou</td>
<td>A Critical Path Approach to Analyzing Parallelism of Algorithmic Variants</td>
<td>133</td>
</tr>
<tr>
<td>Ren-Cang Li</td>
<td>Accurate Solutions of M-Matrix Algebraic Riccati Equations</td>
<td>135</td>
</tr>
<tr>
<td>Xiaoye S. Li</td>
<td>Towards an Optimal Parallel Approximate Sparse Factorization Algorithm Using Hierarchically Semi-separable Structures</td>
<td>137</td>
</tr>
<tr>
<td>Jörg Liesen</td>
<td>On the Convergence of GMRES for a Convection-diffusion Model Problem</td>
<td>139</td>
</tr>
<tr>
<td>Lek-Heng Lim</td>
<td>Fundamental Difficulties of Numerical Multilinear Algebra</td>
<td>141</td>
</tr>
<tr>
<td>Lijing Lin</td>
<td>A Schur–Padé Algorithm for Fractional Powers of a Matrix</td>
<td>142</td>
</tr>
<tr>
<td>Author</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>------------------------------</td>
<td>-----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>James Nagy</td>
<td>A Computational Approach for Large Scale Nonlinear Least Squares Problems</td>
<td>168</td>
</tr>
<tr>
<td>Yuji Nakatsukasa</td>
<td>Optimizing Halley’s Iteration for Computing the Matrix Polar Decomposition</td>
<td>170</td>
</tr>
<tr>
<td>Esmond G. Ng</td>
<td>A Combinatorial Problem in Sparse Orthogonal Factorization</td>
<td>172</td>
</tr>
<tr>
<td>Dianne P. O’Leary</td>
<td>Euclidean Distance Matrix Completion Problems and Protein Structure Determination</td>
<td>174</td>
</tr>
<tr>
<td>Ivan Oseledets</td>
<td>TENSOR TRAIN Decomposition and Its Applications</td>
<td>175</td>
</tr>
<tr>
<td>Michael L. Overton</td>
<td>Fast Algorithms for the Approximation of the Pseudospectral Abscissa and Pseudospectral Radius of a Matrix</td>
<td>176</td>
</tr>
<tr>
<td>Chris Paige</td>
<td>A Backward Rounding Error Analysis of the Lanczos Process, and its Implications for Other Krylov Subspace Methods</td>
<td>177</td>
</tr>
<tr>
<td>Ivo Panayotov</td>
<td>Ritz Vectors in the Lanczos Process</td>
<td>179</td>
</tr>
<tr>
<td>Jennifer Pestana</td>
<td>On Choice of Preconditioner for Minimum Residual Methods for Nonsymmetric Matrices</td>
<td>180</td>
</tr>
<tr>
<td>Martin Plešinger</td>
<td>The Total Least Squares Problem with Multiple Right-Hand Sides</td>
<td>182</td>
</tr>
<tr>
<td>Federico Poloni</td>
<td>A Perron Vector-based Iteration for Solving Quadratic Vector Equations</td>
<td>184</td>
</tr>
<tr>
<td>Stefan Ragnarsson</td>
<td>Symmetric Embeddings of Higher-Order Tensors</td>
<td>186</td>
</tr>
<tr>
<td>Alison Ramage</td>
<td>Iterative Solution of Linear Systems in Liquid Crystal Modelling</td>
<td>187</td>
</tr>
<tr>
<td>Tyrone Rees</td>
<td>A Multipreconditioned GMRES Algorithm</td>
<td>188</td>
</tr>
<tr>
<td>Rosemary Renaut</td>
<td>Recycling Krylov Subspaces for Efficient Schwarz Algorithms with Extensions to Solve Regularized Least Squares Problems</td>
<td>189</td>
</tr>
<tr>
<td>Marielba Rojas</td>
<td>New Methods for Least-Norm Regularization</td>
<td>191</td>
</tr>
<tr>
<td>Miro Rozložník</td>
<td>Orthogonalization with a Non-standard Inner Product and Approximate Inverse Preconditioning</td>
<td>193</td>
</tr>
<tr>
<td>Axel Ruhe</td>
<td>Recent Developments of Rational Krylov Algorithms</td>
<td>195</td>
</tr>
<tr>
<td>Author</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------</td>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Hassane Sadok</td>
<td>A New Approach to GMRES Convergence</td>
<td>196</td>
</tr>
<tr>
<td>Berkant Savas</td>
<td>Multilinear Rank-((r_1, r_2, r_3)) Approximation of a Tensor: Optimality Conditions and Perturbation Theory of Local Optima</td>
<td>199</td>
</tr>
<tr>
<td>Christian Schröder</td>
<td>A Jacobi-Davidson Method for Two Real Parameter Nonlinear Eigenvalue Problems arising from Delay Differential Equations</td>
<td>201</td>
</tr>
<tr>
<td>Oded Schwartz</td>
<td>Graph Expansion and Communication Costs of Fast Matrix Multiplication</td>
<td>203</td>
</tr>
<tr>
<td>Jennifer Scott</td>
<td>The Robust and Efficient Partial Factorization of Dense Symmetric Indefinite Matrices</td>
<td>206</td>
</tr>
<tr>
<td>Mili Shah</td>
<td>Calibrating Robotic Vision Systems</td>
<td>208</td>
</tr>
<tr>
<td>Josef Sifuentes</td>
<td>Convergence Theory for a Restarted GMRES Method with Approximate Deflation Preconditioning</td>
<td>210</td>
</tr>
<tr>
<td>Danny C. Sorensen</td>
<td>Linear Methods in Nonlinear Model Order Reduction</td>
<td>214</td>
</tr>
<tr>
<td>Alastair Spence</td>
<td>A New Algorithm for the Computation of the Real Distance to Instability</td>
<td>216</td>
</tr>
<tr>
<td>Suvrit Sra</td>
<td>Large-scale Nonnegative Least-Squares</td>
<td>217</td>
</tr>
<tr>
<td>G.W. Stewart</td>
<td>When Is Twice Enough? (the oblique case)</td>
<td>220</td>
</tr>
<tr>
<td>Martin Stoll</td>
<td>Preconditioning Time-Dependent Optimal Control problems</td>
<td>222</td>
</tr>
<tr>
<td>Zdeněk Strakoš</td>
<td>Matching Moments and Matrix Computations</td>
<td>224</td>
</tr>
<tr>
<td>Xiaobai Sun</td>
<td>A Divide and Conquer Method for the SVD of a Banded Matrix</td>
<td>226</td>
</tr>
<tr>
<td>Brian D. Sutton</td>
<td>Stable Computation of the CS Decomposition: Simultaneous Bidiagonalization</td>
<td>227</td>
</tr>
<tr>
<td>Daniel B. Szyld</td>
<td>Petrov Galerkin view of IDR (and BiCGStab)</td>
<td>228</td>
</tr>
</tbody>
</table>
Sue Thorne

Preconditioners for PDE-constrained Problems with Nonlinear PDEs in the Constraints .............................................................. 229

Petr Tichý

On Best Approximation by Polynomials of Matrices ......................... 231

Françoise Tisseur

A Reliable Algorithm for the Complete Solution of Quadratic Eigenvalue Problems ............................................................... 233

David Titley-Peloquin

Projected Residuals and Backward Error Estimates in LSQR ............... 235

Miroslav Tůma

On the Way Towards Robust Algebraic Preconditioners .................... 237

Elisabeth Ullmann

Iterative Solvers for Stochastic Galerkin Discretizations of the Lognormal Diffusion Problem ...................................................... 239

Martin B. van Gijzen

An Elegant IDR(s) Variant that Efficiently Exploits Bi-orthogonality Properties ................................................................. 240

Sabine Van Huffel

Successful Tensor Decompositions in Clinical Practice ..................... 242

Charles Van Loan

Block Tensor Computation ............................................................. 243

Raf Vandebril

The Interplay of Givens Rotations and the QR-factorization in Inversion, Unitary Similarity Transforms and QR-algorithms ............... 245

Bart Vandereycken

Riemannian and Multilevel Optimization for Rank-constrained Matrix Problems ............................................................. 246

Stephen A. Vavasis

Finding Approximately Rank-one Submatrices with the Nuclear Norm and $\ell_1$-norm ................................................................. 248

Eugene Vecharynski

Preconditioned Locally Minimal Residual Method for Computing Interior Eigenpairs of Symmetric Operators .......................... 250

Vu Hoang Linh

Spectral Intervals for Differential-algebraic Equations and their Numerical Approximation by QR and SVD Algorithms .................. 251

Kees Vuik

Recursively Deflated PCG for Mechanical Problems ....................... 252

Homer F. Walker

Anderson Acceleration for Fixed-Point Iterations ............................ 254

David S. Watkins

The Next Step in the Never-ending Process of Generalizing Francis’s Implicitly-shifted QR Algorithm ............................................. 255
Jianlin Xia
Efficient Structured Solvers and Preconditioners for Large Sparse Linear Systems ........................................... 256

Fei Xue
Efficient Preconditioned Inner Solves for Inexact Rayleigh Quotient Iteration and Their Connections to the Simplified Jacobi-Davidson Method .......................................................... 258

Ichitaro Yamazaki
A Parallel Hybrid Linear Solver for Large-scale Highly-Indefinite Linear Systems of Equations ................................. 259

Chao Yang
Solving Nonlinear Eigenvalue Problems in Electronic Structure Calculation ....................................................... 261

Bo Zhang
A Fourier-Series-Based Kernel-Independent Fast Multipole Method ................................................................. 262

List of Participants 264
Householder Symposium XVIII on Numerical Linear Algebra

June 12-17, 2011
Tahoe City, California

Householder Symposium XVIII on Numerical Linear Algebra is the eighteenth in a series, previously called the Gatlinburg Symposia. The series of conferences is named after its founder Alston S. Householder, one of the pioneers in numerical linear algebra. The 2011 symposium is hosted by Lawrence Berkeley National Laboratory and is in cooperation with SIAM and SIAG/LA.

The Householder Symposia originated in a series of meetings organized by Alston Householder, Director of the Mathematics Division of Oak Ridge National Laboratory and Ford Professor at the University of Tennessee. These international meetings were devoted to matrix computations and linear algebra and were held in Gatlinburg, Tennessee. They had a profound influence on the subject. The last “Gatlinburg” conference held at Gatlinburg was in 1969 on the occasion of Householder’s retirement. At the time, it was decided to continue the meetings but vary the place. Since then meetings have been held at three-year intervals in a variety of venues and the series has been renamed in honor of Alston Householder. Table 1 contains a complete list of the previous symposia.

The meetings, which last for five days, are by invitation only. They are intensive, with plenary talks in the day and special sessions in the evenings. To encourage people to talk about work in progress, no proceedings are published, although extended abstracts are circulated. The response of the participants to the meetings has generally been very enthusiastic.

The conferences are run in tandem by a permanent organizing committee and a local arrangements committee. Although attendance is restricted, anyone - including students - can apply. Selection is made by the organizing committee, generally by ballot.


The meeting is also the occasion for the award of the Householder prize for the best thesis in numerical linear algebra. The term numerical linear algebra is intended to describe those parts of mathematical research that have both linear algebraic aspects and numerical content or implications. Thus, for example, a dissertation concerned with the numerical solution of differential equations or the numerical solution of an optimization problem would be eligible if linear algebra is central to the research contribution. This prize is entirely (and well) supported by contributions solicited at the Symposium banquet. The Householder Award, given every three years, was established at the 1969 Gatlinburg Symposium to recognize the outstanding contributions of Alston S. Householder, 1904-1993, to numerical analysis and linear algebra. Nominations are assessed by an international committee. Table 2 contains a complete list of the previous winners.

---

1Prof. G.W. Stewart, University of Maryland, has an article on Alston S. Householder in SIAM News, Vol. 26, October 1993.
<table>
<thead>
<tr>
<th>Number</th>
<th>Year</th>
<th>Place</th>
<th>Organizers</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>1961</td>
<td>Gatlinburg, U.S.A.</td>
<td>A.S. Householder</td>
</tr>
<tr>
<td>II</td>
<td>1963</td>
<td>Gatlinburg, U.S.A.</td>
<td>A.S. Householder, F.W.J. Olver</td>
</tr>
<tr>
<td>III</td>
<td>1964</td>
<td>Gatlinburg, U.S.A.</td>
<td>A.S. Householder</td>
</tr>
<tr>
<td>IV</td>
<td>1969</td>
<td>Gatlinburg, U.S.A.</td>
<td>A.S. Householder</td>
</tr>
<tr>
<td>V</td>
<td>1972</td>
<td>Los Alamos, U.S.A.</td>
<td>R.S. Varga</td>
</tr>
<tr>
<td>VI</td>
<td>1974</td>
<td>Hopfen am Seem, BRD</td>
<td>F.L. Bauer</td>
</tr>
<tr>
<td>VII</td>
<td>1977</td>
<td>Asilomar, U.S.A.</td>
<td>G.H. Golub</td>
</tr>
<tr>
<td>IX</td>
<td>1984</td>
<td>Waterloo, CANADA</td>
<td>J.A. George</td>
</tr>
<tr>
<td>X</td>
<td>1987</td>
<td>Fairfield Glade, U.S.A.</td>
<td>R.C. Ward, G.W. Stewart</td>
</tr>
<tr>
<td>XI</td>
<td>1990</td>
<td>Tylosand, SWEDEN</td>
<td>A. Björck</td>
</tr>
<tr>
<td>XII</td>
<td>1993</td>
<td>Lake Arrowhead, U.S.A.</td>
<td>T.F. Chan, G.H. Golub</td>
</tr>
<tr>
<td>XIII</td>
<td>1996</td>
<td>Pontresina, SWITZERLAND</td>
<td>W. Gander, M.H. Gutknecht, D.P. O’Leary</td>
</tr>
<tr>
<td>XIV</td>
<td>1999</td>
<td>Whistler, B.C., CANADA</td>
<td>J.M. Varah, G.W. Stewart</td>
</tr>
<tr>
<td>XV</td>
<td>2002</td>
<td>Peebles, SCOTLAND</td>
<td>P. Knight, A. Ramage, A. Wathen, N.J. Higham</td>
</tr>
<tr>
<td>XVI</td>
<td>2005</td>
<td>Seven Springs, U.S.A.</td>
<td>J. Barlow, D. Szyld, H. Zha, C. Van Loan</td>
</tr>
<tr>
<td>XVII</td>
<td>2008</td>
<td>Zeuthen, GERMANY</td>
<td>J. Liesen, V. Mehrmann, R. Nabben, A. Bunse-Gerstner</td>
</tr>
</tbody>
</table>

Table 1: Previous Householder Symposia.
<table>
<thead>
<tr>
<th>Year</th>
<th>Names</th>
</tr>
</thead>
<tbody>
<tr>
<td>1971</td>
<td>F. Robert (Grenoble)</td>
</tr>
<tr>
<td>1974</td>
<td>Ole Hald (New York University)</td>
</tr>
<tr>
<td>1977</td>
<td>Daniel D. Warner (University of California, San Diego)</td>
</tr>
<tr>
<td>1981</td>
<td>E. Marques de Sa’ (Coimbra); Paul Van Dooren (K. U. Leuven) - shared</td>
</tr>
<tr>
<td>1984</td>
<td>Ralph Byers (Cornell University); James M. Demmel (University of California, Berkeley) - shared</td>
</tr>
<tr>
<td>1987</td>
<td>Nicholas J. Higham (University of Manchester)</td>
</tr>
<tr>
<td>1990</td>
<td>Alan Edelman (Massachusetts Institute of Technology) ; Maria Beth Ong (University of Washington) - shared</td>
</tr>
<tr>
<td>1993</td>
<td>Hong-Guo Xu (Fudan University) ; Barry Smith (New York University) - shared</td>
</tr>
<tr>
<td>1996</td>
<td>Ming Gu (Yale University)</td>
</tr>
<tr>
<td>1999</td>
<td>Jörg Liesen (Bielefeld)</td>
</tr>
<tr>
<td>2002</td>
<td>Jing-Rebecca Li (Massachusetts Institute of Technology)</td>
</tr>
<tr>
<td>2005</td>
<td>Jasper van den Eshof (Utrecht University)</td>
</tr>
<tr>
<td>2008</td>
<td>David Bindel (University of California, Berkeley)</td>
</tr>
</tbody>
</table>

Table 2: Previous Householder Award Winners.
Householder Committee

Tony Chan, Hong Kong University of Science and Technology
Alan Edelman, Massachusetts Institute of Technology
Nick Higham, University of Manchester
Ilse Ipsen, North Carolina State University (SIAM Representative)
Volker Mehrmann, Technical University of Berlin
Michael Overton, Courant Institute, New York University (chair)
Yousef Saad, University of Minnesota
Valeria Simoncini, University of Bologna
Zdenek Strakos, Charles University in Prague
Paul Van Dooren, Université catholique de Louvain
Andy Wathen, University of Oxford

Local Organizer

Esmond G. Ng, Lawrence Berkeley National Laboratory

Householder Prize Committee

Michele Benzi, Emory University
James Demmel, University of California, Berkeley (chair)
Howard Elman, University of Maryland
Volker Mehrmann, Technical University of Berlin
Sabine Van Huffel, Catholic University of Leuven
Stephen Vavasis, University of Waterloo
Acknowledgments

We are grateful to the following organizations for their generous financial support and sponsorship of Householder Symposium XVIII.

Householder Student Travel Fund (provided by the book royalties from *Milestones in Matrix Computation: The Selected Works of Gene H. Golub with Commentaries*, edited by Raymond Chan, Chen Greif and Dianne O’Leary (Oxford University Press, 2007), and *G.W. Stewart: Selected Works with Commentaries*, edited by Misha Kilmer and Dianne O’Leary (Birkhauser Boston, 2010))

International Linear Algebra Society (ILAS)

Lawrence Berkeley National Laboratory

MathWorks

Numerical Algorithms Group (NAG)

Oxford University Press

Society for Industrial and Applied Mathematics (SIAM), and SIAM Activity Group on Linear Algebra (SIAG/LA)

The Dr. Richard Carl Dehmel Distinguished Professorship in Engineering, provided by the holder, Prof. James Demmel, University of California, Berkeley

U.S. Department of Energy

U.S. National Science Foundation
Abstracts

(Abstracts are arranged in alphabetical order of presenters.)
An Improved Algorithm for the Matrix Logarithm

Awad H. Al-Mohy, and Nicholas J. Higham

Abstract

The matrix logarithm is an important matrix function playing a key role in applications. \( X \in \mathbb{C}^{n \times n} \) is a logarithm of \( A \in \mathbb{C}^{n \times n} \) if \( e^X = A \). If \( A \) has no eigenvalues on \( \mathbb{R}^- \) (the closed negative real axis) then there is a unique logarithm with eigenvalues having imaginary parts lying in the interval \((-\pi, \pi)\) [3, Thm. 1.31]; this is known as principal logarithm and denoted by \( \log(A) \).

The inverse scaling and squaring method (ISS) originally proposed by Kenney and Laub [4] begins by performing the Schur decomposition \( A = QTQ^* \), computes \( T^{1/2s} \) with an integer \( s \geq 0 \) large enough to bring \( T^{1/2s} \) close to the identity, approximates \( \log(T^{1/2s}) \) by \( r_m(T^{1/2s} - I) \), where \( r_m(x) \) is the \([m/m]\) Padé approximant to the function \( \log(1 + x) \), and finally recovers \( \log(A) \) using

\[
\log(A) = 2^s \log(A^{1/2s}) = 2^s Q \log(T^{1/2s}) Q^*.
\]

We use the \([m/m]\) Padé approximant in the partial fraction form [2]

\[
r_m(x) = \sum_{j=1}^{m} \frac{a_j^{(m)} x}{1 + \beta_j^{(m)} x},
\]

where the \( a_j^{(m)} \) and \( \beta_j^{(m)} \) are the weights and the nodes, respectively, of the \( m \)-point Gauss-Legendre quadrature rule on \([0, 1]\).

In this work the performance of the ISS method is considerably improved by

- introducing new backward error analysis for Padé approximation of the matrix logarithm,
- obtaining sharp bounds for the backward error using techniques introduced in the analysis of Al-Mohy and Higham [1],
- replacing the diagonal and first superdiagonal elements of the matrix \( 2^s r_m(T^{1/2s} - I) \) by highly accurate computed components before undoing the Schur decomposition in (1).

We give some details and begin by the following lemma.

**Lemma 1.** Let \( A \in \mathbb{C}^{n \times n} \) have no eigenvalues on \( \mathbb{R}^- \), and suppose that \( \rho(A) < 1 \). Then

\[
\rho(r_m(A)) \leq \sum_{j=1}^{m} \frac{a_j^{(m)} \rho(A)}{1 - \beta_j^{(m)} \rho(A)}.
\]

Define the matrix function \( h_{2m+1}(A) = e^{r_m(A)} - A - I \) and suppose that \( \rho(A) < 0.91 \). We find practically using (3) that \( \rho(r_m(A)) < \pi \) for \( 1 \leq m \leq 16 \), which implies that \( \log(e^{r_m(A)}) = r_m(A) \) [3, Prob. 1.39]. Thus we can write

\[
r_m(A) = \log(I + A + h_{2m+1}(A)) = \log(I + A + \Delta A),
\]

\[7\]
where $\Delta A = h_{2m+1}(A)$. Hence the matrix function $h_{2m+1}(A)$ represents the backward error resulting from the approximation of $\log(I + A)$ by $r_m(A)$ and has the power series expansion

$$h_{2m+1}(A) = \sum_{k=2m+1}^{\infty} c_k A^k.$$ 

We bound the backward error using [1, Thm. 4.2(a)] as follows:

$$\frac{\|\Delta A\|}{\|A\|} \leq \sum_{k=2m+1}^{\infty} |c_k| \alpha_p(A)^{k-1},$$

where $\alpha_p(A) = \max(d_p, d_p + 1)$, $d_p = \|A^p\|^{1/p}$, with $p$ arbitrary subject to $2m + 1 \geq p(p - 1)$. $d_p$ can be estimated cheaply using matrix norm estimators. Let $\theta_m = \max\{ \theta : \sum_{k=2m+1}^{\infty} |c_k| \theta^{k-1} \leq \text{tol} \}$, where tol is a given tolerance.

Now we outline our improved algorithm for computing $\log(A)$.

- Evaluate the Schur decomposition $A = QTQ^*$.
- Find the $s$ and $m$ that minimize the cost of the algorithm subject to $\alpha_p(T^{1/2^s} - I) \leq \theta_m$.
- Evaluate $X := 2^s r_m(T^{1/2^s} - I)$.
- Replace $\text{diag}(X)$ by $\log(\text{diag}(T))$ and the elements of the first superdiagonal of $X$ by those given by [3, Eq. (11.28)].
- Undo the effect of unitary transformation: $\log(A) \approx QXQ^*$.

The state of the art algorithms described in [3, Chap. 11] for $\log(A)$ determine $m$ and $s$ such that $\|W\| \leq \theta_m$, where $W = T^{1/2^s} - I$, where $\theta_m$ is obtained from forward error considerations. Our approach defines $\theta_m$ via backward error analysis and determines $m$ and $s$ so that $\alpha_p(W) \leq \theta_m$. Since $\alpha_p(W) \leq \|W\|$ always holds, and $\alpha_p(W) \ll \|W\|$ is possible for nonnormal $W$, our new algorithm is potentially much faster. Our numerical experiments show this and confirm that the new algorithm delivers better accuracy.

References


Interpolation on Matrix Manifolds of Reduced-Order Models and Application to On-Line Aeroelastic Predictions

David Amsallem

Abstract

Projection-based Reduced-Order Models (ROMs) have seen a growing interest in the computational engineering community, as their lower dimensionality implies reduced computational costs. Unfortunately, routine analysis involves parameter variations and most if not all ROMs lack robustness with respect to parameter changes. Therefore, performing computations with ROMs calls for reconstructing a new ROM every time a new configuration is considered. However, such a reconstruction can be a computationally intensive process since the high-fidelity model is involved.

Together, these two issues underline the need for a new strategy for adapting pre-computed ROMs to new sets of physical or modeling parameters. In this work, databases of reduced-order information associated to fast interpolation-based techniques are considered. ROMs and their corresponding reduced-order bases are quantities that typically belong to nonlinear, matrix manifolds. As such, classical interpolation methods fail, as they are not able to enforce the constraints characterizing those manifolds.

The first part of this work consists in identifying the quantities of interest to be interpolated as well as their associated constraints and then designing a suitable interpolation method enforcing the constraints. In particular, it is shown that the Grassmann manifold is the manifold of interest when interpolating reduced-order bases. Applications to the fast aeroelastic prediction of the behavior of two full aircraft configurations (F-16 Block 40 and F-18/A) using databases of fluid reduced-order bases are then presented.

The second main contribution of this work is the development of a novel consistent interpolation procedure of parameterized linear reduced-order operators. These matrices are computed by projection of full-size matrices onto their respective reduced-order bases. As a result, they are expressed in a different set of coordinates for each instance of the parameters. It is shown that applying appropriate congruence transformations enable expressing these operators into consistent sets of coordinates. These transformations are computed as the solution of small-size orthogonal Procrustes problems which possess a closed-form analytical solution. After transformation of the reduced-order operators into these consistent sets of coordinates, interpolation on the appropriate manifold can occur. This two-step procedure is shown to greatly reduce the computational cost for aeroelastic predictions while retaining good accuracy. Unlike the method of interpolation on the Grassmann manifold, this method operates solely on small-size linear operators. As a result, this whole procedure was implemented on a smartphone and shown to deliver predictions in real-time.

The proposed methods enable test operation calls for new, “last minute” flight configurations and thus paves the way for on-line, routine usage of reduced-order modeling including during flight testing.
Often the behavior of dynamical systems depends on parameters, such as material properties, geometric characteristics or varying boundary conditions. Parametrized systems are abundant and can be found in connection with MEM (microelectromechanical) Systems, electronic chip design, interconnect modeling, etc.

The preservation of these parameters as symbolic variables in the reduced models is important for applications. This leads to parametrized model reduction. The reduced order models should produce responses which are close to those of the original system for the whole range of variation of the design parameters.

The dominant approach to parametric systems is currently interpolatory model reduction, see for instance see [4, 3, 2], and references therein. The reduction problem is thereby approached by matching the input/output behavior at specified values of the frequency and of the parameters. Hence, the question of effective frequency and parameter point selection arises.

In some applications, the designer specifies important parameter sets to be used in model reduction. However, the task becomes much harder if one only has the information as the range of parameter space without any knowledge of what parameter sets might be important.

The approach that we propose aims at bypassing this problem. The transfer function of a linear system depending on the parameters $\tau_i$, $i = 1, \cdots, k$, can be considered as a $(k+1)$-variable rational function, the variables being the complex frequency $s$ and the parameters $\tau_i$. Therefore the study of such systems can be reduced to the study of rational functions depending on several variables.

In the several-variable rational function setting, our approach consists in collecting input/output (e.g. frequency response) measurements for some appropriate range of frequencies and appropriate range of parameter values. Then using the new methodology which we will refer to as Loewner matrix framework, one can construct models which are reduced with respect both to frequency as well as to the parameter dependence. The key is that in contrast to the existing approaches, larger amounts of data than necessary are collected and the essential underlying system structure is extracted appropriately.

This presentation concerns the case $k = 1$ (one-parameter systems) and is the first step towards a treatment of the general parametrized model reduction problem as described above. The basic advantage of this approach is that it provides the user with a systematic methodology for constructing low complexity two-variable rational interpolants.

The talk will follow the development in [1]. We will first present our main tool, the Loewner matrix. In particular, after a brief overview of the single-variable case [5], the two-variable extension will be developed. It will then be shown that the null-space of an appropriately defined Loewner matrix contains all the information necessary for constructing low-degree interpolants from given samples.

Next, we will discuss generalized state-space realizations of these interpolants. The problem of minimality of realizations in the two-variable case being open, we show how to obtain low dimensional realizations which are valid for general rational functions (including polynomials).

Up to this point it has been assumed that the interpolation points are distinct. The next part of the presentation will be dedicated to two-variable interpolation with multiplicities. Particular
attention is paid to the case when all data is provided at the origin (which is equivalent to partial realization of two-variable rational functions).

Besides the material in [1], we will also present as yet unpublished results concerning the use of the two-variable Loewner matrix for the reduction of systems depending on one parameter. The tool for achieving this is a structured singular value decomposition involving of the associated Loewner matrix.

One of the major issues arising is that the quantities involved have complexity which is multiplicative in the complexity $n$ with respect to $s$, and $m$ with respect to the parameter $\tau_1$. One of our main results is that these objects can be determined with computational complexity which is essentially additive, with respect to $n$ and $m$. This makes the Loewner framework computationally feasible.

The presentation will end with examples both academic and of practical relevance, illustrating the Loewner-matrix approach to parametrized model reduction.

References


Generalized Golub-Kahan Bidiagonalization and Stopping Criteria

Mario Arioli,
STFC Rutherford Appleton Laboratory (UK)

Abstract

Let $M \in \mathbb{R}^{m \times m}$ and $N \in \mathbb{R}^{n \times n}$ be symmetric positive definite matrices, and let $A \in \mathbb{R}^{m \times n}$ be a full rank matrix. By using $M$ and $N$, we define the following Hilbert spaces

$$M = \{ v \in \mathbb{R}^m; \|u\|_M^2 = v^T M v \}, \quad N = \{ q \in \mathbb{R}^n; \|q\|_N^2 = q^T N q \}$$

and their dual spaces

$$M' = \{ w \in \mathbb{R}^m; \|w\|_{M^{-1}}^2 = w^T M^{-1} w \}, \quad N' = \{ y \in \mathbb{R}^n; \|y\|_{N^{-1}}^2 = y^T N^{-1} y \}.$$ 

Using the previous notation, the matrix $A$ is an operator mapping from $N$ into $M$. In particular, for each fixed $q \in N$ we also have

$$\langle v, Aq \rangle_{M, M'} = v^T Aq, \quad Aq \in \mathcal{L}(M) \ \forall q \in N.$$ 

The adjoint operator $A^*$ of $A$ can be defined as

$$\langle A^* g, f \rangle_{N', N} = f^T A^T g, \quad A^T g \in \mathcal{L}(N) \ \forall g \in M,$$

and it is linked to the transpose of $A$. Given $q \in M$ and $v \in N$, the critical points for the functional

$$\frac{v^T Aq}{\|q\|_N \|v\|_M}$$

are the "generalized singular values and singular vectors" of $A$. Indeed the saddle-point conditions for (1) are

$$\begin{align*}
Aq_i &= \sigma_i M v_i, \\
A^T v_i &= \sigma_i N q_i
\end{align*}$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$. If we perform a change of variables using $M^{1/2}$ and $N^{1/2}$ the generalized singular values of $A$ are the standard singular values of $\tilde{A} = M^{-1/2} A N^{-1/2}$.

The generalized singular vectors $q_i$ and $v_i$, $i = 1, \ldots, n$ are the transformation by $M^{-1/2}$ and $N^{-1/2}$ of the left and right standard singular vector of $\tilde{A}$ respectively.

We point out that the necessary and sufficient conditions, based on the $\inf$-$\sup$ condition [3] that guarantee both existence and unicity of the solution, and the stability of saddle-point variational problems, are equivalent to the conditions that the generalized singular values $\sigma_i$ of $A$ are in the interval $(a, b)$ with $0 < a < b$ and $a$ and $b$ independent of the dimensions $n$ and $m$. This also implies that the generalized condition number $\kappa(A) = \frac{\sigma_1}{\sigma_n}$ is independent of $n$ and $m$.

In the following, we analyse the use of the generalized Golub-Kahan bidiagonalization algorithm [4], where the scalar products will be based on $M$ and $N$, in solving the problem

$$\min_{A^T u = b} \|u\|_M^2,$$
where $M$ is a nonsingular symmetric and positive definite matrix.

Several problems can be reduced to case (2). The general problem

$$\min_{A^Tw = r} \frac{1}{2} w^T W w - g^T w,$$

where the matrix $W$ is positive semidefinite and $\ker(W) \cap \ker(A^T) = 0$ can be reformulated as (2) by choosing

$$M = W + \nu AN^{-1} A^T, \quad u = w - M^{-1} g, \quad b = r - A^T M^{-1} g.$$

If $W$ is nonsingular then we can choose $\nu = 0$.

The proposed algorithm can be used for the solution of augmented systems that give the optimality conditions for (2), and, in particular, we propose a novel estimator of the error similar to that proposed by Hestenes-Stiefel for the conjugate gradient algorithm. This estimator gives a lower bound for the error, and can be used as a reliable stopping criterion for the whole process. We also propose an upper bound of the error based on Gauss-Radau quadrature. Furthermore, we show how we can transform and optimally precondition augmented systems rising from the mixed finite-element approximation of differential problems [1] so that we achieve a rate of convergence independent of the mesh size.

Finally, we will describe how it is possible to adapt the bidiagonalization process in order to have conjugate gradient-like algorithms tuned for the solution of symmetric quasi-definite systems [2, 5] such as

$$\begin{bmatrix} M & A \\ A^T & -N \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} q \\ b \end{bmatrix}. \quad (3)$$

These particular systems arise frequently in optimization and in the regularization of linear systems of PDE from a mixed finite-element approximation.

**References**


Randomized Algorithms in Numerical Linear Algebra: From Theory to Practice

Haim Avron and Sivan Toledo

Abstract

Randomization is arguably the most exciting and innovative idea to have hit numerical linear algebra in a long time. Several new randomized algorithms have been proposed and explored in the past decade. Some forms of randomization have been used for decades in linear algebra. For example, the starting vectors in Lanczos algorithms are always random. But recent research led to new uses of randomization: random mixing and random sampling, which can be combined to form random projections. These ideas have been explored theoretically and have found use in some specialized applications (e.g., data mining), but they have had little influence so far on mainstream numerical linear algebra.

We will show that if the new techniques are fused with older, well established techniques, they can be as stable as traditional algorithms, but much faster. More specifically, we will describe a least-square solver for dense highly overdetermined systems that achieves residuals similar to those of direct factorization based state-of-the-art solvers (LAPACK), outperforms LAPACK by large factors, and scales significantly better than any QR-based solver. The solver uses a traditional iterative solver to solve the system, but the preconditioner is built using a randomized algorithm. This fusion of randomization and preconditioned iterative solvers is what makes our solver fast and reliable.

The talk is based on two recent papers. The first paper, titled "Randomized Algorithms for Estimating the Trace of an Implicit Symmetric Positive Semi-definite Matrix" [3], was recently accepted to the Journal of the ACM. This paper rigorously analyzes the convergence of randomized trace estimators.

Starting at 1989, several algorithms have been proposed for estimating the trace of a matrix by \( \frac{1}{M} \sum_{i=1}^{M} z_i^T A z_i \), where the \( z_i \) are random vectors [6]; different estimators use different distributions for the \( z_i \)'s, all of which lead to \( E(\frac{1}{M} \sum_{i=1}^{M} z_i^T A z_i) = \text{trace}(A) \). These algorithms are useful in applications in which there is no explicit representation of \( A \) but rather an efficient method compute \( z^T A z \) given \( z \). For example, in lattice Quantum Chromodynamics, one often needs to compute the trace of a function of a large matrix, \( \text{trace}(f(A)) \). Explicitly computing \( f(A) \) for large matrices is not practical, but computing the bilinear form \( x^T f(A) x \) for an arbitrary \( x \) is feasible [5, 4]. Other examples include the regularized solution of least-squares problems using the Generalized Cross-Validation approach (see [6]), estimating \( \| A^{-1} \|_F \) and computing the number of triangles in a graph [1].

Previous results only analyze the variance of the different estimators. In contrast, we analyze the number of samples \( M \) required to guarantee that with probability at least \( 1 - \delta \), the relative error in the estimate is at most \( \epsilon \). We argue that such bounds are much more useful in applications than the variance. We also found that these bounds rank the estimators differently than the variance; this suggests that minimum-variance estimators may not be the best. We found that the number of samples for the best estimator \( O(\epsilon^{-2} \log(1/\delta)) \). We argue that such bounds, although typical in randomized algorithms, are not good enough for numerical linear algebra applications because typically they require very high accuracy.

The second paper, titled "Blendenpik: Supercharging LAPACK’s Least-Squares Solver" [2] was recently published in SIAM Journal on Scientific Computing. In this paper we describe Blendenpik,
a new randomized least squares solver for dense systems. Through careful engineering, extensive analysis and experimentation, we have been able to use randomization to build an algorithm is that is faster than a state-of-the-art linear solver. Blendenpik beats LAPACK’s direct dense least-squares solver by a large margin on essentially any dense tall matrix. Blendenpik is slower than LAPACK on tiny matrices, nearly square ones, and on some sparse matrices. But on a huge range of matrices of reasonable sizes Blendenpik is much faster than LAPACK. In fact, on large matrices, Blendenpik is about four times faster than LAPACK. We believe that these results show the potential of random-sampling algorithms, and suggest that random-projection algorithms should be incorporated into future numerical linear algebra libraries.


References


Progress in Linear and Nonlinear Eigensolvers

Zhaojun Bai

Abstract

Algebraic linear and nonlinear eigenvalue problems appear in a wide variety of forms in real-life applications. In this talk, I will first describe the origins and challenges of some eigenvalue problems arising from our recent collaborative work in electronic structure calculations and solar energy research, and then survey a couple of the solution algorithms that have been recently proposed and discuss work-in-progress.

Specifically, we will first revisit the popular symmetric Lanczos method and examined performance issues of the method with respect to the choice of projection subspace dimension. This is motivated by the fact that on modern computer systems, communication (memory access) costs have come to exceed arithmetic costs by orders of magnitude, and the gap is only growing over time. In order to achieve optimal performance, it is necessary to dynamically select the dimension of projection subspace that balances memory (I/O) costs and convergence rate. The re-design of linear algebra algorithms that reduce or even provably minimize communication costs has attracted a lot of attention lately. We will report our work on how to identify a set of tuning parameters to determine the subspace dimension on-the-fly in the thick-restart Lanczos method of Simon and Wu to minimize the overall run-time.\textsuperscript{1} Then we will describe our work-in-progress of extending the idea to other linear eigensolvers and applications to large-scale random phase approximation eigenvalue problems and calculations of optical absorption spectra of nanoparticles arising from solar energy research.\textsuperscript{2}

In the second half of this talk, we will focus on nonlinear eigenvalue problems (NEPs). Different from those NEPs that are nonlinear in the eigenvalue parameter such as polynomial and rational NEPs, we will consider the forms of NEPs that are nonlinear with respect to eigenvectors, formally written as “$H(\mathbf{X})\mathbf{X} = S\mathbf{X}\Lambda$”. A major source of this form of NEPs is from solving the Kohn-Sham equation for electronic structure calculations, although it also appears in image reconstruction and feature extraction and other applications. A common approach to the NEPs is via a triple-loop iteration, oftenly referred to as a self-consistent field (SCF) method. We will describe an integrated approach to analyze the relationship between loops and to properly set error control parameters for each loop to optimize the overall convergence rate.\textsuperscript{3}

\textsuperscript{2}G. Zimanyi, Z. Bai, G. Galli, S. Kauzlarich, D. Larsen and D. Paul, \textit{NSF SOLAR Collaborative project on “Multiple exciton generation and charge extraction in all-inorganic nanostructured solar cells”}, work-in-progress.
\textsuperscript{3}Z. Bai, J. Pask, N. Sukumar, Y. Nakatsukasa and Y. Cai. \textit{Project on “Parallel solvers for eigenproblems in quantum mechanics”}, work-in-progress.
Avoiding Communication using Successive Band Reduction

Grey Ballard, Jim Demmel, Ioana Dumitriu, and Nicholas Knight

Abstract

The running time of an algorithm depends on two factors: the number of floating point operations executed (arithmetic) and the amount of data moved (communication) between levels of a memory hierarchy by a sequential algorithm or over a network connecting processors by a parallel algorithm. The simplest metrics of communication are the total number of words moved (the bandwidth cost) and the total number of messages containing these words (the latency cost). On current hardware the cost of moving a single word already greatly exceeds the cost of one arithmetic operation, and technology trends indicate that this processor-memory gap is growing exponentially over time. This motivates our efforts to devise algorithms that minimize communication, sometimes even at the price of doing more arithmetic.

In this talk we present communication-avoiding algorithms for three problems:

1. reducing a symmetric matrix to tridiagonal form using orthogonal similarity transformations,
2. reducing a nonsymmetric matrix to bidiagonal form using two-sided orthogonal transformations,
3. computing the eigenvectors of a matrix in Schur form.

The new algorithms do $O(n^3)$ arithmetic operations, are numerically stable, and communicate asymptotically less than previous such algorithms.

In fact, these algorithms attain known communication lower bounds that apply to many $O(n^3)$ algorithms in dense linear algebra. In the sequential case, when the $n$-by-$n$ input matrix does not fit in fast memory of size $M$, the number of words moved between fast (small) and slow (large) memory is $\Omega(n^3/\sqrt{M})$ and the number of messages moved is $\Omega(n^3/M^{3/2})$. In the case of $P$ parallel processors, where each processor has room in memory for $1/P$-th of the input matrix, the number of words moved between one processor and the others is $\Omega(n^2/\sqrt{P})$ and the number of messages sent is $\Omega(\sqrt{P})$. These lower bounds were originally proven for sequential [10] and parallel [2] matrix multiplication, and extended to many other linear algebra algorithms in [3], including the three algorithms we discuss here.

In contrast, conventional algorithms for these problems, including those currently implemented in (Sca)LAPACK, perform asymptotically more communication than the lower bounds require. In the sequential case, they communicate $O(n^3)$ words, and in the parallel case, they communicate $O(n)$ messages.

The new algorithms for first two problems are variants on the conventional reduction to tridiagonal (for the symmetric eigenvalue problem) or bidiagonal form (for the SVD), after which a symmetric tridiagonal eigensolver (respectively bidiagonal SVD) running in $O(n^2)$ time is used to complete the problem. The reduction itself is split into two phases: reducing the full matrix to banded form and reducing the banded matrix to tridiagonal/bidiagonal form. The first phase utilizes a communication-avoiding QR decomposition algorithm introduced in [4], and the second phase consists of a process known as successive band reduction discussed at length in [5].

Recent work [6] demonstrated that, with the right variant of successive band reduction, this two-step approach attains the communication lower bound for the sequential bandwidth cost proved in [3].
This approach performs more arithmetic than the conventional reduction algorithms, although the increase is a lower-order term if only eigenvalues (or singular values) are requested. If eigenvectors (or both left and right singular vectors) are also desired, then the arithmetic cost of the reduction is increased by $2\times$ for large matrices (when one column of the input matrix is too large to fit into fast memory) and $2.6\times$ for moderately sized matrices (when the matrix is too large to fit into fast memory, but one or more columns do fit into fast memory).

The third problem listed above concerns the computation of the eigenvectors of a nonsymmetric matrix from its Schur decomposition. We present in [6] methods for both computing the Schur decomposition of a nonsymmetric matrix and finding its eigenvectors. Ioana Dumitriu's extended abstract (submitted to the Householder Symposium) discusses a set of communication-avoiding algorithms for computing Schur form. In this talk, we will assume the matrix to be in Schur form, and we present both parallel and sequential algorithms which reorganize the computation of the eigenvectors and minimize both the bandwidth and latency costs. These algorithms perform exactly the same number of arithmetic operations as conventional algorithms.

Our ongoing research addresses open problems related to successive band reduction, both in algorithm design and implementation. We plan to adapt the sequential algorithm and the underlying banded matrix layout to minimize the latency cost (in addition to minimizing the bandwidth cost).

We will also analyze the communication and arithmetic complexity of existing parallel successive band reduction algorithms; our preliminary findings are that these algorithms do not attain the communication lower bounds. We plan to adapt and improve them to minimize communication, keeping the cost of any extra arithmetic as low as possible.

We will implement our sequential and parallel algorithms, the latter targeted at both distributed and shared memory machines. Ultimately, we will demonstrate the importance of avoiding communication and the effectiveness of our new algorithms with performance comparisons against existing implementations.

References


Positivity Preserving Discretizations for Differential-Algebraic Systems

Ann-Kristin Baum

Abstract

In the simulation of dynamical processes in economy, social sciences, biology or chemistry, the analyzed values typically represent 'real' quantities like the amount of money, goods or individuals or the concentration or density of chemical and biological species. From the mathematical point of view, this leads to positive systems, i.e., differential or difference equations with the property that every non-negative initial value yields a non-negative solution for all times and thus reflects the material character of the considered variable.

Examples of positive systems are e.g. Leontief- and Leslie-Models, semi-discretized advection-diffusion-equations arising for instance in atmospheric chemistry or chemotaxis problems as well as compartmental models used in biology, pharmacokinetics or epidemiology, see e.g. [1, 2].

Beside the positivity, such processes often have to satisfy additional constraints resulting from limitation of resources, conservation laws or balance conditions and which extend the differential system by accessory algebraic equations.

Solving these differential-algebraic equations (DAE) numerically, we thus encounter two problems: we must satisfy the algebraic constraints and meet the positivity condition in order to obtain physical meaningful results.

In my talk, I will explain how the existing results of unconstrained positive systems, see e.g. [1, 3], can be generalized to DAEs and what needs to be done to circumvent the sometimes severe stepsize restrictions for the considered Runge-Kutta- and Multistep-Methods.

The essential point in this analysis will be a capable tool giving access to the constraining subspace to which every solution of a DAE is bounded. For linear time-invariant problems $E \dot{x} = Ax + f$ with $E, A \in \mathbb{R}^{n \times n}$ and $x, f \in \mathbb{R}^n$, this subspace corresponds to a constant hyperplane that can be identified under certain regularity conditions with the kernel of $E$. In this case, we obtain a projection onto this plane with the Drazin inverse of the matrix $E$ and are thus able to consider the dynamic and the algebraic components of the DAE and its discretization separately. This way, we can nicely extend the existing results of unconstrained linear systems, like for example the famous result of Bouley and Crouzeix [4], by considering rational functions operating on generalized inverses. For the algebraic components, we will use the consistency of the applied method to derive conditions and stepsize estimates for a positive preserving approximation.

With this unified approach, I will conclude my talk by pointing out the chances to relax these stepsize restrictions for particular problems. Using the theory of matrix valued functions, I will explain how the special case of cyclic matrices with a multiple eigenvalue enforces the strict stepsize restrictions in the classical positivity results and how they can be softened, if we exploit the structure of the given problem more precisely. For a single matrix, i.e., for unconstrained problems, this will lead to quantifying the 'positive impact' of the off-diagonal entries, where for constrained systems we have to filter out the relevant, positivity providing entries of the matrix pair. Again, we will do this with the help of the projection onto the constraining subspace.
References


Model Reduction of Hamiltonian Systems in Variational Data Assimilation

Christopher Beattie\textsuperscript{*}, Timothy Campbell\textsuperscript{**}, Serkan Gugercin\textsuperscript{*}, and Caleb Magruder\textsuperscript{*}

Abstract

In the face of limited computational resources and rapid turn-around times, the need for timely and accurate simulations of processes in the ocean and atmosphere creates an acute demand for effective strategies to reduce the cost of evaluation of the related hydrodynamic models without losing the level of accuracy that is necessary for mission requirements. The context where these costs become especially forbidding lies in determining physically plausible, appropriate model initializations and forcings that are capable of bringing simulations into coherence with available observations. Since observations cannot be taken at every location in the model domain, the distributed inputs that a model receives as well as its initial state can deviate dramatically from what occurs in the real atmosphere and ocean. This remains as one of the primary sources of uncertainty and forecast errors — and data assimilation techniques for the systematic adjustment of dynamical system forcing and initial conditions have become the primary response to these challenges. Model reduction methods in turn offer significant potential for generating compact representations of dynamical systems that can capture essential features of the original model central for the effective application of data assimilation techniques.

One approach to variational data assimilation, the representer method, has as its principal computational burden the repeated solution of a large-order Hamiltonian boundary value problem. To illustrate this, consider a nonlinear evolution problem \( \dot{\eta} = f(\eta) \) with a nominal “background” trajectory, \( \eta_b(t) \). The background trajectory constitutes the best prediction of model behaviour available based on prior observations and hypothesized forcings for the forecast interval. For simplicity, we consider only the prospect of small errors in dynamics and ignore initial conditions as a possible source of error. Suppose that subsequent observations are made on the actual system trajectory, \( \eta(t) \) at times \( t_1, t_2, ..., t_m \):

\[
y_1 = \ell(\eta(t_1)), \quad y_2 = \ell(\eta(t_2)), \quad y_3 = \ell(\eta(t_3)), \quad ..., \quad y_m = \ell(\eta(t_m))
\]

using (say) a single linear functional \( \ell(\eta) = c^T \eta \). Generally the observations won’t match the predictions:

\[
\tilde{y}_1^b = \ell(\eta_b(t_1)), \quad \tilde{y}_2^b = \ell(\eta_b(t_2)), \quad \tilde{y}_3^b = \ell(\eta_b(t_3)), \quad ..., \quad \text{with } \gamma_k = y_k - \tilde{y}_k^b \neq 0.
\]

We wish to adjust the hypothesized system dynamics,

\[
\dot{\eta} = f(\eta) \quad \Rightarrow \quad \dot{\eta} = f(\eta) + \rho(t)
\]

by including a residual forcing term, \( \rho(t) \), that will be used to accommodate observations. The new observations \( \{y_1, y_2, ..., y_m\} \) generally will not completely determine \( \rho(t) \), nor would we wish them to since observations inevitably will contain some error. So, we seek residual forcings that create a minimal perturbation to the background trajectory in an appropriate sense while bringing the modified trajectory into greater coherence with the new observations, also in an appropriate sense. This may be formulated as follows - suppose that our background trajectory has been specified on

\footnotesize
\textsuperscript{*}Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061
\textsuperscript{**}Ocean Dynamics and Prediction Branch, Naval Research Laboratory 7320, Stennis Space Center, MS 39529
$(-\infty, \infty)$ and define a perturbation to the background trajectory as $\mathbf{z} = \eta(t) - \eta_b(t)$. Now, our task is

Find $\rho(t) \in L^2(\mathbb{R})$ such that

$$\int_{-\infty}^{\infty} \mathbf{z}(t)^T \mathbf{Wz}(t) + \rho(t)^T \mathbf{M} \rho(t) \, dt + \sum_{k=1}^{m} w_k (\gamma_k - \ell(\mathbf{z}(t_k)))^2$$

is minimized

for $\dot{\mathbf{z}} = \mathbf{Jz} + \rho(t)$ with $\mathbf{J} = \left[ \frac{\partial f_i}{\partial \eta_j} \right] \eta_b$ (the Jacobian matrix for $f$)

Here $\mathbf{W}$ and $\mathbf{M}$ are positive definite matrices that establish in effect penalties for deviation from the background trajectory and $\{w_1, w_2, \ldots, w_m\}$ are positive weights that establish a penalty for deviation from the observations. $\mathbf{J}$, the Jacobian matrix for $f$ evaluated on the background trajectory, is typically time-varying.

The representer method framework leads in a natural way to a (big) Hamiltonian system:

$$\dot{\mathbf{x}} = \begin{bmatrix} \mathbf{J} & \mathbf{M}^{-1} \\ \mathbf{W} & -\mathbf{J}^T \end{bmatrix} \mathbf{x} + \begin{pmatrix} 0 \\ \mathbf{c} \end{pmatrix} u(t)$$

such that $\mathbf{x}(t)$ is $L^2(\mathbb{R})$

$$y(t) = [\mathbf{c}^T \, 0] \, \mathbf{x}(t)$$

defined for inputs $u(t) \in L^2(\mathbb{R})$. This is a (noncausal) dynamical system that must be solved repeatedly for each observation time $t_1, t_2, \ldots, t_m$ with an additional boundary value constraint given by $\mathbf{x}(t_k^-) - \mathbf{x}(t_k^+) = \begin{pmatrix} 0 \\ \mathbf{c} \end{pmatrix}$, for $k = 1, \ldots, m$. This is a daunting task if many observations are to be made.

Ideally, we want to precompute as much as possible in order minimize forecast delays that occur due to processing time in the period after observations have been made. More generally, both the observation values $y_1, y_2, \ldots, y_m$ and the exact observation times $t_1, t_2, \ldots, t_m$ may be unknown during this preprocessing phase.

We consider the utility of model reduction approaches in finding compact representations of (1). Typically, model reduction methodology is applicable only to stable or nearly stable systems and is most highly developed for linear time invariant dynamical systems. Aside from linearity, the system (1) enjoys none of these traits. Nonetheless, interpolatory model reduction methods offer a degree of flexibility and versatility that permit a systematic approach to finding effective reduced order surrogates for (1). We have recently extended the optimal $H_2$ interpolation method for model reduction to arbitrary unstable systems that are recast as boundary value problems (instead of initial value problems). We have also extended these methods to time-varying systems. We discuss how these techniques may be deployed in the context of (1) and how they may then be applied to variational data assimilation problems arising in coastal ocean hydrodynamic models.
Rational Krylov Subspaces for Nonlinear Model Reduction

Peter Benner, Tobias Breiten

Abstract

We discuss Krylov-subspace based model reduction techniques for nonlinear control systems. Since reduction procedures of existent approaches like TPWL and POD methods [3, 7] require simulation of the original system and are therefore dependent on the chosen input function, models that are subject to variable excitations might not be sufficiently approximated. We will overcome this problem by generalizing Krylov-subspace methods known from linear systems to a more general class of bilinear and quadratic-bilinear systems, respectively. As has recently been shown, a lot of nonlinear dynamics can be represented by the latter systems. We will explain advantages and disadvantages of the different approaches and discuss the choice of reasonable interpolation points with regard to optimal approximation results.

To be specific, we consider nonlinear control-affine systems

\[ \dot{x}(t) = f(x(t)) + Bu(t), \quad y(t) = Cx(t), \quad x(0) = x_0, \]  

with \( f : \mathbb{R}^n \to \mathbb{R}^n \) nonlinear, \( B \in \mathbb{R}^{n \times m}, C \in \mathbb{R}^{m \times n}, x \in \mathbb{R}^n, u \in \mathbb{R}^m, y \in \mathbb{R}^m \). We now want to construct a reduced system

\[ \dot{\hat{x}}(t) = \hat{f}(\hat{x}(t)) + \hat{B}u(t), \quad \hat{y}(t) = \hat{C}\hat{x}(t), \quad \hat{x}(0) = \hat{x}_0, \]

with \( \hat{f} : \mathbb{R}^\hat{n} \to \mathbb{R}^\hat{n}, \hat{B} \in \mathbb{R}^{\hat{n} \times m}, \hat{C} \in \mathbb{R}^{m \times \hat{n}}, \hat{x} \in \mathbb{R}^{\hat{n}}, \hat{y} \in \mathbb{R}^m, \hat{n} \ll n \) s.t. \( \hat{y} \approx y \) for all admissible \( u \).

The model reduction approaches discussed here are based on representing (1) as quadratic-bilinear control systems (QBDAE), given as follows:

\[ E\dot{x}(t) = A_0x(t) + A_1x(t) \otimes x(t) + A_2(I_m \otimes x(t))u(t) + Bu(t), \quad y(t) = Cx(t), \quad x(0) = x_0, \]

where \( A_0, E \in \mathbb{R}^{N \times N}, A_1 \in \mathbb{R}^{N \times Nm}, A_2 \in \mathbb{R}^{N \times N^2}, B, C^T \in \mathbb{R}^{N \times m} \). It has recently been shown in [5] that many systems with nonlinearities composed of uni-variable rational, exponential, logarithmic, trigonometric or root functions, respectively, can be transformed into a system of QBDAEs as in (2) of dimension \( N > n \) by iteratively taking Lie derivatives and adding algebraic equations. Note that even if the original system is an ODE as in (1), the corresponding QBDAE may have a singular \( E \)-matrix. In contrast to the QBDAE representation which is exact, Carleman bilinearization is used to approximate a nonlinear system of the form (1). Truncation of the resulting approximation leads to a bilinear system representation which corresponds to \( E = I_N \) and \( A_1 = 0 \) in (2) [9].

Projection-based model reduction methods for QBDAE and bilinear systems now consist in finding appropriate \( r \)-dimensional (\( r \ll n \)) subspaces \( V \) of \( \mathbb{R}^N \) on which the system is projected to obtain a reduced-order model as

\[ \hat{E} = W^TEV, \quad \hat{A}_0 = W^TA_0V, \quad \hat{A}_1 = W^TA_1V \otimes V, \quad \hat{A}_2 = W^TA_2(I_m \otimes V), \quad \hat{B} = W^TB, \quad \hat{C}^T = W^TC, \]

where \( \text{range}(V) = V \) and the oblique projection of the state-space is given by \( P := VW^T \) with \( W \in \mathbb{R}^{n \times r} \) representing a complimentary subspace.

In the talk, we will touch on the following issues, some of which are also discussed in [2]:

- Using variational analysis and Volterra series, it is possible to derive input/output relations and transfer functions similar to the linear case. In particular, the choice of transfer functions in the QBDAE case is non-unique, and we discuss the advantages of choosing symmetric transfer functions.
• Interpolating these transfer functions yields moment-matching model reduction methods. We will review earlier approaches, e.g., by [1, 5, 4], which suggest (rational) Krylov subspace methods for computing these rational interpolants. As in these approaches, only single-point interpolation is considered and the complimentary subspace is chosen as $W = V$ (resulting in an orthogonal projection method), we will in particular discuss multi-point tangential interpolation and two-sided projection methods, i.e., $W \neq V$.

• A particularly important issue for a good approximation is the choice of interpolation points. In the linear case, optimality conditions for $H_2$-norm approximation have led to the IRKA algorithm suggested in [6], where the interpolation points are chosen so that the rational interpolant satisfies these optimality conditions. Similar $H_2$-optimality conditions for bilinear systems are suggested in [10]. We will show that an appropriate choice of Krylov subspaces together with an IRKA-like iteration can be used to compute rational interpolants that satisfy these $H_2$-optimality conditions and that preserve the bilinear model structure.

We will also discuss the widely open problem of choosing interpolation data for the QBDAE reduced-order models.

Numerical examples from circuit simulation, chemical engineering, computational neurosciences, and flow control will be used to illustrate the abilities of the new nonlinear model reduction methods and in particular highlight the capability of yielding good simulation results for varying input functions using the same reduced-order model.

References


Abstract

In this talk I will describe certain linear algebra problems arising in the modeling of complex networks. As shown recently by E. Estrada, D. Higham, N. Hatano and others, quantitative methods of network analysis naturally lead to large-scale computations for functions of matrices associated with graphs, including adjacency matrices and graph Laplacians. Typical functions include the matrix exponential, the resolvent, and hyperbolic functions. After introducing the basic notions of subgraph centrality, Estrada index, betweenness, communicability, etc., I will describe numerical techniques for obtaining bounds and estimates for such quantities. Relationships between the matrix functions and spectral properties of the underlying graph, such as the existence of gaps, will be investigated. If time allows, some extensions to directed and weighted graphs will also be discussed. Finally, I will mention some open problems in this area.

This is joint work with Paola Boito (University of Limoges) and Christine Clymko (Emory University).
A Modular and Systematic Approach to Stability Analysis

*Paolo Bientinesi and Robert van de Geijn*

**Abstract**

Numerical stability analysis related to dense linear algebra operations continues to be an important topic in numerical analysis. It is part of almost any introductory and advanced text on numerical linear algebra and earned one of its pioneers, J.H. Wilkinson, a Turing Award in 1970. We introduce a systematic methodology to derive error analyses of numerical dense linear algebra algorithms.

Parlett observes that “One of the major difficulties in a practical analysis is that of description. An ounce of analysis follows a pound of preparation.” This suggests the desirability of a standardization of the description of error analyses. Higham echoes this sentiment: “Two reasons why rounding error analysis can be hard to understand are that, first, there is no standard notation and, second, error analyses are often cluttered with re-derivations of standard results.” His statement also suggests that an inventory of results, to be used in subsequent analyses, is desirable. It is our belief that textbooks and papers alike focus on the error results and their proofs rather than on exposing a methodology that can be used by practitioners and students to obtain error results in a systematic way.

We approach the topic from a new perspective: We are interested in the methodology for deriving results as much as, if not more than, the results themselves. Our goal is to identify notation and a procedure for error analyses that can, in principle, be made mechanical (automatic). As part of the FLAME project, these objectives have already been accomplished in the derivation of algorithms for linear algebra operations: We identified notation and exposed a systematic procedure that was reproducible by a computer algebra system. We show now that the same notation and procedure can be extended to equally systematically (although not yet mechanically) derive stability analyses.

We highlight here the salient characteristics of the methodology.

- The analysis of a target algorithm builds modularly, composing the analyses for the operations that appear in the body of the algorithm.

- The approach is goal-oriented: One starts formulating a possible error result, and then performs a sequence of steps to validate it. In the process, results for subproblems will be exposed. The methodology can then be applied to derive that subresult.

- Multiple analyses for the same operation can be derived and catalogued for future reuse. Such an inventory of (sub)results is used to construct the analysis of a target operation.

We do not claim that our approach is different from what an expert does as he/she analyzes an algorithm. Our claim is that this methodology is systematic enough that 1) experts and non-experts alike can derive new analyses, and 2) it is within reach to develop a computer algebra system to make the derivation process mechanical.
A Matrix Factorization for Computer Network Tomography

David Bindel

Abstract

Computer network tomography systems reconstruct properties of links in a network from measurements of routing paths that traverse those links. In many cases, the measured path properties are sums of link properties; for example, the latency of a routing path is the sum of the latencies of the links. We write this relationship between path and link properties as $Gx = b$, where $x$ is a vector of link properties, $b$ is a vector of path properties, and $G$ is a matrix whose rows are indicator vectors for the links used on various paths, i.e.

$$G_{ij} = \begin{cases} 1, & \text{when path } i \text{ traverses link } j, \\ 0, & \text{otherwise.} \end{cases}$$

The matrix $G$ is typically large, sparse, and highly rank deficient. In previous work, we developed network tomography algorithms that require that we choose a small set of rows of $G$ that span the row space of $G$. In this setting, $G$ typically has several thousand columns, several tens of thousands of rows, and a rank of several hundred to a few thousand. Thus, the cost of the factorization used to pick a row basis for $G$ is a scalability problem.

In this talk, we describe a new factorization of $G$ as a product of sparse rectangular matrices:

$$G = \left[ I_{T_p} \right] \left[ I_{T_{p-1}} \right] \ldots \left[ I_{T_1} \right] \bar{G},$$

where $G \in \{0, 1\}^{m_0 \times n}$ consists of a few rows of $G$ and $T_j \in \{0, 1, -1\}^{(m_j - m_{j-1}) \times m_{j-1}}$. The method works by finding linear dependencies among paths that intersect at a common router. To find dependencies among the paths crossing a given router $R$, we first write each path as the sum of a source segment (a path segment going from a source to $R$) and a destination segment (a path segment leading from $R$ to a destination). In matrix terms, if $G_R$ is a submatrix of $G$ corresponding to all paths that go through router $R$, then we can write each row of $G$ as $s + d$ for some source segment indicator $s$ and destination segment indicator $d$. Note that if $s_1 + d_1, d_1 + s_2, s_2 + d_3, \ldots, s_{k-1} + d_k, d_k + s_1$ are rows in $G_R$, then we have

$$(s_1 + d_1) - (d_1 + s_2) + (s_2 + d_3) - \ldots + (s_{k-1} + d_k) - (d_k + s_1) = 0.$$ 

If we identify rows of $G_R$ with edges in a bipartite graph whose nodes represent source and destination segments, such a linear dependency corresponds to cycle in the graph. Moreover, the edges of a spanning forest in this graph correspond to rows of $G_R$ that form a spanning set for the row space of $G_R$. Based on this graph-theoretic interpretation of submatrices of $G$, we describe a fast greedy algorithm to factor $G$ and a compact representation of the factorization. We illustrate the effectiveness of our approach through experiments with real and artificial computer networks.
Commute Times for a Directed Graph using an Asymmetric Laplacian

Daniel Boley, Gyan Ranjan, and Zhi-Li Zhang

Abstract

The spectral analysis of undirected graphs has been studied extensively [1, 5, 7, 9, 10, 12, 15, 16, 17, 18, 19, 20], but fewer papers exist discussing the spectral analysis of directed graphs (digraphs) [2, 4, 6, 21]. In particular, the relationship between expected first transit/hitting times and round-trip commute times in a random walk, on the one hand, and spectral properties of the underlying graph on the other, has been studied mainly for undirected graphs. In this paper, we show that the round-trip commute times are closely related to certain asymmetric “Laplacian” matrices for strongly connected directed graphs in ways analogous to those known for undirected graphs. We show that one can approximate a strongly connected digraph by a related weighted undirected graph which shares some of the properties of the original digraph (e.g. connectivity, stationary probabilities), while only approximately inheriting others (e.g. first transit/hitting times and node centrality). This has applications in domains with asymmetric connections, such as wireless packet switching networks with low-powered units where link asymmetry is a widely observed phenomenon.

A directed graph, or digraph, \( G = (V, E) \), is a collection of vertices (or nodes) \( i \in V = \{1, \ldots, n\} \) and directed edges \( i \rightarrow j \in E \). One can assign weights to each directed edge, thereby making it a so-called weighted digraph, or else use a common edge weight of 1 to obtain an unweighted digraph. A directed graph \( G \) is called strongly connected or a strong digraph if there is a path \( i = \ell_0 \rightarrow \ell_1 \rightarrow \cdots \rightarrow \ell_{\kappa-1} \rightarrow \ell_\kappa = j \) for any pair of nodes \( i, j \), where each link \( \ell_{i-1} \rightarrow \ell_i, i = 1, \ldots, \kappa \), is an edge in the graph. In this paper, we focus entirely on strongly connected directed graphs.

A random walk over a graph can be modeled by a Markov chain with probability transition matrix \( P = D^{-1}A \), where \( D = \text{Diag}(d) = \text{Diag}(A \cdot 1) \) is the diagonal matrix of vertex out-degrees, \( A \) is the adjacency matrix, and \( 1 \) denotes the vector of all ones. In a strong digraph, every node has at least one out-going edge and one incoming edge. The associated vector of stationary probabilities is denoted by \( \pi \) and satisfies \( \pi^T P = \pi^T \) and \( \pi^T 1 = 1 \). If the graph is strongly connected, the associated Markov chain is irreducible, and all the entries of \( \pi \) are strictly positive by Perron-Frobenius theory [11, 13]. If the graph were undirected, the associated Markov chain would be reversible, and the vector of stationary probabilities would be a scalar multiple of the vector of vertex degrees: \( \pi = d/(d^T 1) \), where the denominator would be called the volume of the graph. Unfortunately, this relationship does not necessarily hold for digraphs. We denote by \( \Pi = \text{Diag}(\pi) \), the diagonal matrix of stationary probabilities, which is non-singular if the graph is strongly connected.

In this work, we examine scaled “Laplacian” matrices, not necessarily symmetric and denoted by \( L = \Pi(I - P) \) and \( L^d = \Pi^{-\frac{1}{2}}L\Pi^{-\frac{1}{2}} = \Pi^{-\frac{1}{2}}(I - P)\Pi^{-\frac{1}{2}} \), together with their Moore-Penrose pseudo-inverses \( M = L^+, M^d = (L^d)^+ \), which are defined for a strongly connected directed graph or a strong digraph. Even though most of the derivations mimic known derivations for undirected graphs, not everything carries over from the world of undirected graphs to that of their directed counterparts. For example, the concept of “volume” of a graph and the metaphor of resistances of an electrical network [3, 8, 14] do not play the obvious central role in the derivations for directed graphs as they do for undirected graphs.

We show the following for strongly connected directed graphs:

a. The average hitting times and round-trip commute times can be expressed in terms of the pseudo-inverses of the asymmetric Laplacians.
b. The commute time is the square of an Euclidean distance measure for the vertices of a strongly connected directed graph, and \((M+M^T)/2\) is the related Gram matrix of inner products.

c. There is a close relationship between the so-called Fundamental Matrix and the pseudo-inverse \(M\) of the asymmetric Laplacian \((L)\).

d. The commute times for a directed graph can be bounded in terms of the stationary probabilities and the eigenvalues of the diagonally scaled symmetrized graph Laplacian.

e. Simply symmetrizing the Laplacian as previously proposed [6, 21] yields a weighted undirected graph with the same links (now all bidirectional) and the same \(\pi\), but guaranteed higher average commute times. The average one-way hitting times can be arbitrarily higher.

Sample Results. The expected hitting \(H\), and commute times \(C\) satisfy the following

- \(H(i,j) = \frac{m^d_{ij}}{\pi_j} - \frac{m^d_{ij}}{\sqrt{\pi_i \pi_j}}\), where \(m_{ij}\), \(m^d_{ij}\) are the \(i,j\)-th entries of \(M = L^+, M^d = (L^d)^+\), resp.

- \(C(i,j) = m_{jj} + m_{ii} - m_{ij} = \frac{m^d_{ij}}{\pi_j} + \frac{m^d_{ij}}{\pi_i} - \frac{m^d_{ij} + m^d_{ji}}{\sqrt{\pi_i \pi_j}}\).

- \(H(i,j) \leq tr[M^d]/\pi_j\).

- \(\max\left\{\frac{1}{\pi_i}, \frac{1}{\pi_j}\right\} \leq C(i,j) \leq tr[M^d] \cdot \left(\frac{1}{\pi_j} + \frac{1}{\pi_i}\right)\). The left-hand inequality can be tight.

- \(C(i,j) \leq C^s(i,j)\) (the commute times for the symmetrized graph).

References

Moving Eigenvalues with Structured Perturbations, Crawford Number and ε-pseudospectra

Shreemayee Bora, and Ravi Srivastava

Abstract

It is well known that the eigenvalues of a complex Hamiltonian matrix $H$ are symmetrically placed with respect to the imaginary axis as they occur in pairs $(\lambda, -\lambda)$. It was established in [1] that if all the eigenvalues of $H$ are purely imaginary, then the smallest Hamiltonian perturbation $\Delta H$ with respect to the 2 norm such that any arbitrarily small Hamiltonian perturbation to $H + \Delta H$ results in eigenvalues that are not purely imaginary for the perturbed matrix is given by the smallest value of $\epsilon$ for which components of the Hamiltonian $\epsilon$-pseudospectrum of $H$ containing eigenvalues with opposite sign characteristics coalesce. A parallel question may be asked for eigenvalues of Hermitian pencils $(A, B)$ (for which $A^* = A$ and $B^* = B$) where the role of the imaginary axis for Hamiltonian matrices is played by the real axis. In other words, given such a pencil $(A, B)$ with all eigenvalues real, what is the smallest perturbation $(\Delta A, \Delta B)$ with respect to some prespecified norm such that the perturbed pencil $(A + \Delta A, B + \Delta B)$ is Hermitian and has the property that any arbitrarily small structure preserving perturbation results in a pencil with eigenvalues that are not real?

We find that in this case besides the sign characteristics of eigenvalues of $(A, B)$, the solution also depends on the algebraic multiplicity of the smallest singular value $\sigma_n(A - z_0 B)$ of $A - z_0 B$ where $z_0$ is a point of coalescence of two components of the Hamiltonian $\epsilon_0$-pseudospectrum of the pencil $(A, B)$ and $\epsilon_0$ is the smallest value of $\epsilon$ for which such coalescence takes place. To be precise we observe the following:

**Theorem 1.** Let $(A, B)$ be a complex matrix pencil of size $n$ with $A^* = A, B^* = B$ and all eigenvalues real and distinct. Let $\Lambda(A, B)$ denote the spectrum of $(A, B)$ and consider the Hermitian $\epsilon$-pseudospectrum $\Lambda^H_\epsilon(A, B) := \cup \{ \Lambda(A + E, B + F) : E^* = E, F^* = F, \| (E, F) \| < \epsilon \}$ of $(A, B)$ with respect to the norm $\| (A, B) \| := \sqrt{\| A \|^2 + \| B \|^2}$. Let $\epsilon_0$ be the smallest value of $\epsilon$ for which $\Lambda^H_\epsilon(A, B)$ has less than $n$ components and $z_0$ be a point of coalescence of some two components of $\Lambda^H_\epsilon(A, B)$. Also suppose that $\sigma_n(A - z_0 B)$ is a simple singular value of $A - z_0 B$. Then $z_0$ is a defective eigenvalue of the pencil $(A + \Delta A, B + \Delta B)$ where

$$\Delta A := -\frac{\sigma_n(A - z_0 B)}{1 + z_0^2} v_0 v_0^*, \quad \text{and} \quad \Delta B := z_0 \frac{\sigma_n(A - z_0 B)}{1 + z_0^2} v_0 v_0^*,$$

$v_0$ being a right singular vector of $A - z_0 B$ corresponding to $\sigma_n(A - z_0 B)$.

If $\sigma_n(A - z_0 B)$ is multiple and the components of $\Lambda^H_\epsilon(A, B)$ coalescing at $z_0$ contain eigenvalues of opposite sign characteristic, then $z_0$ is either a defective eigenvalue or an eigenvalue of mixed sign characteristic of $(A + \Delta A, B + \Delta B)$ where

$$\Delta A := -\frac{\sigma_n(A - z_0 B)}{1 + z_0^2} [v_0, \hat{v}_0][v_0, \hat{v}_0]^* \quad \text{and} \quad \Delta B := z_0 \frac{\sigma_n(A - z_0 B)}{1 + z_0^2} [v_0, \hat{v}_0][v_0, \hat{v}_0]^*,$$

$v_0, \hat{v}_0$ being appropriately chosen right singular vectors of $A - z_0 B$ corresponding to $\sigma_n(A - z_0 B)$ such that $v_0^* \hat{v}_0 = 0$.

It can be shown that the assumption about coalescing components containing eigenvalues with opposite sign characteristics for the case when $\sigma_n(A - z_0 I)$ is multiple is necessary. Since the
perturbed Hermitian pencils \((A + \Delta A, B + \Delta B)\) as constructed in Theorem 1 have a defective eigenvalue or an eigenvalue with mixed sign characteristic, the canonical forms for such pencils under congruence [2] imply that there exist Hermitian perturbations \((E, F)\) of these pencils which may be chosen to have arbitrarily small norms such that the resulting pencil \((A + \Delta A + E, B + \Delta B + F)\) has eigenvalues that are not real. Moreover if the pencil \((A, B)\) of Theorem 1 is definite as well, then the Crawford number \(\gamma(A, B) = \epsilon_0\) thus leading to a method for computing the Crawford number from the \(\epsilon\)-pseudospectrum of \((A, B)\).

Theorem 1 may also be extended to the case of polynomials \(P(\lambda) := \sum_{k=0}^{n} \lambda^k A_k\), where each \(A_k, k = 1, 2, \ldots, n\) is an \(n \times n\) Hermitian matrix and all eigenvalues of \(P(\lambda)\) are real. Whenever all eigenvalues of \(P(\lambda)\) are of definite type, this will lead to pseudospectra based techniques for finding minimal Hermitian perturbations that result in polynomials outside a particular class (like, for instance hyperbolic, or definite) to which \(P(\lambda)\) belongs.

References


Computing a Nearest Correlation Matrix with Factor Structure

Rüdiger Borsdorf\textsuperscript{1}, Nicholas J. Higham\textsuperscript{2}, and Marcos Raydan\textsuperscript{3}

Abstract

A correlation matrix $C \in \mathbb{R}^{n \times n}$ is a matrix with three defining properties: symmetry, positive semidefiniteness, and unit diagonal. The corresponding nearness problem has lately received considerable attention and a number of efficient algorithms have been developed to solve this problem: Higham’s alternating projection method [3], the inexact primal-dual path following algorithms by Toh, Tüüncü, and Todd [6], Malick’s quasi Newton algorithm [5], and our inexact semismooth preconditioned Newton method proposed in [2] which was originally developed by Qi and Sun. The latter has quadratic convergence and is according to our knowledge the method of choice.

One reason for the increased interest is the large number of practical applications involving statistical modelling. In practice these algorithms are used when it is required to adjust an approximate, empirically obtained correlation matrix such that it satisfies the three defining properties. A valid correlation matrix is often a necessary condition to analyse the obtained data with further methodologies and to prevent it from breaking down. Lack of definiteness in the empirically obtained matrix can result from missing data or asynchronous data which, in case of financial modelling, may be due to a company being formed or ceasing to trade during the period of interest or markets in different regions trading at different times. Furthermore, stress testing may require individual correlations to be artificially adjusted and can cause the matrix to violate the properties of a correlation matrix.

In this talk we extend the nearest correlation matrix problem to include $k$ factor structure $C(X) = \text{diag}(I - XX^T) + XX^T$ where $I$ is the identity matrix and $X$ is a matrix with $n$ rows and $k$ columns and usually $k \ll n$. Our starting point in this research was an algorithm by Andersen, Sidenius and Basu [4]. According to Google Scholar, this paper has currently been cited 234 times showing the importance of the considered problem. Furthermore, we have found an considerable number of references, listed in [1], mainly in the area of finance, where this structure occurs and where the nearness problem is of possible interest.

This factor structure primarily arises when a factor model $\xi = X\eta + D\varepsilon$ is used to describe multivariate time series or collateralized debt obligations. In this model $\eta \in \mathbb{R}^k$ and $\xi \in \mathbb{R}^n$ are vectors of independent random variables having zero mean and unit variance, with $\eta$ and $\varepsilon$ independent of each other. The matrix $X \in \mathbb{R}^{n \times k}$ and $D \in \mathbb{R}^{n \times n}$ is diagonal. In the case of modelling debt obligations $\xi$ can, e.g., model the equity returns of $n$ different companies of a portfolio where $\eta$ describes $k$ factors influencing all companies, in contrast to the elements of $\varepsilon$ having only an effect on the equity of the corresponding company. With this model one tries to capture the complex behaviour of a portfolio of, for example, thousands of equities, by looking at the major factors driving this behaviour.

The factor model is a well established tool in financial modelling and plays an important role in the factor analysis. This gives us further support to think that our proposed nearness algorithm is of major interest in practical applications. As our analysis gives also a reliable tool to check how well a given matrix is approximated by a $k$ factor matrix this can also give support for financial analysts to validate whether a factor model is appropriate for their analysis.

\textsuperscript{1}School of Mathematics, University of Manchester, Manchester, M13 9PL, UK (borsdorf@maths.man.ac.uk).
\textsuperscript{2}School of Mathematics, University of Manchester, Manchester, M13 9PL, UK (higham@maths.man.ac.uk).
\textsuperscript{3}Departamento de Cómputo Científico y Estadística, Universidad Simón Bolívar, Ap. 89000, Caracas 1080-A, Venezuela (mraydan@usb.ve).
NAG, the Numerical Algorithms Group, agreed with the need for algorithms to solve optimization problems arising in the financial area and included in the first instance our nearest correlation matrix code in their library. After receiving frequent requests and a high demand for that algorithm they subsequently released a parallelised version with surprisingly good scaling. As NAG are also convinced that our new algorithm for solving the $k$ factor problem is of interest to their customers they also implemented recently the algorithm that arose from our analysis of this topic. The corresponding routine will be available in the next release and plans to parallelise this algorithm are being made.

In this talk we first analyse the properties of the matrices $C(X)$ and obtain an explicit formula for the rank in the one factor case. In addition, we look at the corresponding nearness problem measured in the Frobenius norm where we impose natural nonlinear constraints on the elements of $X$ ensuring that the resulting matrix is a correlation matrix. For a special one parameter case an explicit solution is obtained. For the general problem, which unlike the nearest correlation matrix problem is nonconvex, we obtain the gradient and the Hessian of the objective function so that first and second order iterative algorithms can be applied. Also, when $k = 1$ an instructive result on the positive definiteness of the Hessian is derived.

We investigate several numerical methods for solving the nearness problem: the alternating directions method; a principal factor method used by Andersen, Sidenius and Basu which we show is equivalent to the alternating projection method, projecting onto a convex set and a nonconvex set in turn and hence lacks convergence results; the spectral projected gradient method (SPGM) of Birgin, Martínez, and Raydan; and Newton and sequential quadratic programming methods. The methods differ in whether or not they can take account of the nonlinear constraints and in their convergence properties. Since all methods are iterative methods we look at the effect of different starting matrices, including a new rank one starting matrix, on the performance. Our numerical experiments show that the performance of the methods depends strongly on the problem, but that SPGM is the clear winner. In addition, we demonstrate empirically for this method how the performance and the optimal objective function value vary when $k$ is increased.

References


Residual, Restarting and Richardson Iteration for the Matrix Exponential

Mike A. Botchev

Abstract

A well-known problem in the computation of matrix functions is the lack of a clear notion of a residual for many matrix functions. Although it is possible to define a residual for some matrix functions such as the inverse or the square root, for many important matrix functions including the matrix exponential, no natural notion for residuals seems to exist.

Let \( A \in \mathbb{R}^{n \times n} \) be a matrix such that \( \frac{1}{2}(A + A^T) \) is positive definite and \( b \in \mathbb{R}^n \) be a given vector. Assume \( y_k \) is a vector computed after \( k \) iterations of some numerical process, such that \( y_k \approx y \equiv \exp(-A)b \). With some abuse of notation, denote by \( y(t) \) a vector function that solves the initial value problem

\[
y'(t) = -Ay(t), \quad y(0) = b.
\]

Clearly, the vector \( y \) is a value of the vector function \( y(t) \) at \( t = 1 \): \( y = y(1) \). Since \( y \) is essentially a value of the exact solution of (1), the equation \( y' = -Ay \) seems to be a natural candidate to define a residual for \( y_k \approx y \). Indeed, let \( y_k(t) \) be a vector function that solves (1) approximately and such that \( y_k(1) = y_k \). We then define the residual \( r_k \) as a value of the function \( r_k(t) \) such that

\[
r_k(t) = -Ay_k(t) - y_k'(t).
\]

In the table below this residual for the matrix exponential is compared against the conventional residual that arises in linear system solution. Both problems (computing the matrix exponential and linear system solution) are instances of computing a matrix function action \( f(A)b \).

<table>
<thead>
<tr>
<th>Residuals for linear system solution and the matrix exponential</th>
<th>( f(x) )</th>
<th>( 1/x )</th>
<th>( \exp(-x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>exact solution ( y )</td>
<td>( y = A^{-1}b )</td>
<td>( \text{define } y(t) = \exp(-tA)b, ) ( \text{set } y := y(1) )</td>
<td></td>
</tr>
<tr>
<td>residual equation</td>
<td>( Ay = b )</td>
<td>{ ( y'(t) = -Ay(t) ), ( y(0) = b ) }</td>
<td></td>
</tr>
<tr>
<td>residual for ( y_k \approx y )</td>
<td>( r_k = b - Ay_k )</td>
<td>( r_k(t) = -Ay_k(t) - y_k'(t) )</td>
<td></td>
</tr>
<tr>
<td>mapping error ( e_k \to \text{residual } r_k )</td>
<td>( r_k = Ae_k )</td>
<td>{ ( r_k(t) = e_k'(t) + Ae_k(t) ), ( e_k(0) = 0 ) }</td>
<td></td>
</tr>
<tr>
<td>perturbed problem (backward stability)</td>
<td>( Ay_k = b - r_k )</td>
<td>{ ( y_k'(t) = -Ay_k(t) - r_k(t) ), ( y_k(0) = b ) }</td>
<td></td>
</tr>
</tbody>
</table>

In this talk we show that the residual for the matrix exponential as defined by (2) can cheaply be computed within standard methods for evaluating the matrix exponential. We describe how to do this for the Chebyshev series expansion and for the Krylov subspace methods, including restarted and shift-and-invert versions. Controlling the residual yields a reliable stopping criterion for this type of methods. Furthermore, new residual-based methods can be constructed. In particular, we consider a Richardson iteration for the matrix exponential. We demonstrate that Richardson iteration can be seen as an efficient way to restart Krylov subspace methods for computing \( \exp(-A)b \).

We expect that a preprint on this topic will be available at [www.math.utwente.nl/~botchevma/](http://www.math.utwente.nl/~botchevma/) since mid November.
Theoretical Foundations for Computation of Eigendecompositions of Third-Order Tensors

Karen Braman and Misha Kilmer

Abstract

Tensors and multilinear algebra are becoming an increasingly important area of research. One of several ways to generalize the familiar tools of linear algebra is to consider third-order tensors as linear mappings in the same way that matrices represent linear transformations between vector spaces. This point of view is facilitated by the definition of a multiplication between third-order tensors introduced in (Kilmer, Marin and Perrone, 2008 and Kilmer and Martin, 2010) and is further investigated in (Braman, 2010). This talk will present recent extensions of these fundamental ideas with an aim toward understanding tensor Krylov subspaces and their use in the computation of tensor eigendecompositions.

Given third-order tensors $A \in \mathbb{R}^{\ell \times p \times n}$ and $B \in \mathbb{R}^{p \times m \times n}$, the $t$-product $A \ast B$ is again a third-order tensor, this time with dimensions $\ell \times m \times n$. For any tensor $A$, let $\vec{A}_i$ denote the $i$th lateral slice of $A$. Note that, for $A$ and $B$ with appropriate dimensions, $A \ast B = [A \ast \vec{B}_1, A \ast \vec{B}_2, \ldots, A \ast \vec{B}_m]$ (1) That is, the $t$-product of $A$ and $B$ can be viewed as $A$ acting on the lateral slices of $B$. With this in mind, it is natural to consider a tensor as a mapping on some space of matrices where the matrices are oriented into the page. In order to provide the most familiar generalization of linear algebra, the space of matrices should have the property that every linear transformation on that space can be represented by tensor multiplication. The usual vector space of $m \times n$ matrices over the reals does not meet this condition and so we require a different collection of scalars. Consider elements of $\mathbb{R}^n$ as third-order tensors oriented into the page, that is, as elements of $\mathbb{R}^{1 \times 1 \times n}$. Then the $t$-product is a commutative operation on this set and, together with component-wise addition, $\mathbb{R}^{1 \times 1 \times n}$ is a commutative ring with unity (Braman, 2010). We denote this ring by $\mathbb{K}_n$ and refer to its elements as tubal-scalars.

Tubal-scalars can be seen as the basic building blocks of both the matrices (the elements of the space) and the tensors (the transformations on the space). The laterally oriented matrices in $\mathbb{R}^{m \times 1 \times n}$ are vectors of tubal-scalars and the tensors in $\mathbb{R}^{\ell \times m \times n}$ are matrices of tubal-scalars. To emphasize this new viewpoint we denote those sets as $\mathbb{K}_m^m$ and $\mathbb{K}_n^{\ell \times m}$, respectively. $\mathbb{K}_n^m$, though not a vector space (because $\mathbb{K}_n$ is not a field), is a free module over $\mathbb{K}_n$ and thus has many of the familiar features of a vector space such as existence of a basis, linear independence and subspaces.

The motivation for Kilmer, Martin and Perrone’s definition of the $t$-product was to produce a tensor analogue of the singular value decomposition. Indeed, they have demonstrated the existence and computation of a $t$-SVD and also of other tensor factorizations. In particular, given a square-faced tensor $A \in \mathbb{K}_n^{m \times m}$, under certain conditions, there exist $\mathcal{X}, \mathcal{D} \in \mathbb{K}_n^{m \times m}$ such that $\mathcal{X}$ is invertible with respect to the $t$-product, $\mathcal{D}$ is $f$-diagonal (every frontal slice is a diagonal matrix) and $A = \mathcal{X}^{-1} \ast \mathcal{D} \ast \mathcal{X}$. That is, $A \ast \mathcal{X} = \mathcal{X} \ast \mathcal{D}$ and so, by (1), $A \ast \vec{X}_j = \vec{X}_j \ast d_{jj}$ for $j = 1, \ldots m$. Thus each $(\vec{X}_j, d_{jj})$ is an eigenpair of $A$. So for tensors, rather than eigenvalues and eigenvectors, we have eigentuples $d_{jj}$ with corresponding eigenmatrices $\vec{X}_j$. In the context of $\mathbb{K}_n^m$, this decomposition should provide information regarding the action of $A$ on $\mathbb{K}_n^m$ and regarding $A$’s invariant subspaces.
Now that we have the appropriate space over which to view the action of a tensor, a natural next step is to consider computation of the eigendecompositions by tensor variations of the Power Method and Krylov subspace methods. At first glance it would seem that everything is in place to be able to produce a tensor version of the Power Method. However, further investigation shows that convergence of the method requires a form of normalization. This is due to the nature of the t-product and its definition in terms of block circulant matrices.

To establish the normalization, we first define a bilinear form on $\mathbb{K}^m$ by $\langle \vec{X}, \vec{Y} \rangle := \vec{X}^T \ast \vec{Y}$, where the transpose is the tensor transpose defined in Kilmer, Martin and Perrone (2008). Of course, this form does not map to the reals, and so in the traditional sense does not define an inner product. In fact, $\langle \vec{X}, \vec{X} \rangle$ is a tubal-scalar and, since it is possible for $\langle \vec{X}, \vec{X} \rangle$ to have negative entries, the standard notion of positivity of the bilinear form does not hold. On the other hand, we can show that the $(1,1,1)$ entry in that tubal-scalar, denoted $\langle \vec{X}, \vec{X} \rangle^{(1)}$, is the square of the Frobenius norm of $\vec{X}$. Thus, it is zero only when $\vec{X}$ is 0, and otherwise, it must be non-negative. So, given $\vec{X} \neq 0 \in \mathbb{K}^n$, we define the length of $\vec{X}$ as $\|\vec{X}\| := \frac{\|\langle \vec{X}, \vec{X} \rangle\|_F}{\sqrt{\langle \vec{X}, \vec{X} \rangle^{(1)}}} = \frac{\|\vec{X}\|_F}{\|\vec{X}\|_F}$. Note that $\vec{X}$ can only have unit length if $\langle \vec{X}, \vec{X} \rangle = e_1$ and that when $n = 1$, so that $\vec{X}$ is in $\mathbb{K}^m$, this definition coincides with the $2$-norm of that vector.

The Power Method then proceeds as expected: beginning with a randomly chosen $\vec{X} \in \mathbb{K}^m$, multiply $A \ast \vec{X}$, normalize the result, and repeat. In most cases, this process converges to the eigenmatrix of $A$ corresponding to the eigentuple of largest magnitude.

eigenpairs. As in the matrix case, given $A \in \mathbb{K}^{m \times m}$ and $\vec{X} \in \mathbb{K}^m$, we define the $j$th Krylov subspace associated with $A$ and $\vec{X}$ as

$$\mathcal{K}_j(A, \vec{X}) := \text{t-span} \left\{ \vec{X}, A \ast \vec{X}, A^2 \ast \vec{X}, \ldots, A^{j-1} \ast \vec{X} \right\}$$

In this definition $t$-span refers to the collection of all linear combinations whose coefficients are tubal-scalars. Again, paralleling the matrix case, the hope is that $\mathcal{K}_j(A, \vec{X})$, with $j$ relatively small, provides an appropriate space in which to search for eigenmatrices of $A$.

While much work remains to be done to understand iterative eigendecomposition methods for tensors, a good first step is to establish a relationship between tensor Krylov subspaces and tensors with a Hessenberg-like form similar to Watkins’ presentation of the matrix case described in The Matrix Eigenvalue Problem (2007). A complication that distinguishes the tensor case from the matrix case is the existence of non-invertible tubal-scalars and in particular, the possibility of their presence as sub-diagonal elements. Thus, we propose the following definition: a square-faced tensor $H \in \mathbb{K}_n^{m \times m}$ is in proper $t$-Hessenberg form if every sub-diagonal tubal-scalar is invertible. With this definition we can establish the following theorem:

**Theorem 1.** Suppose $A, H$ and $G \in \mathbb{K}_n^{m \times m}$ with $G$ invertible, $H = G^{-1} \ast A \ast G$ and let $\vec{X} = \vec{G}_1 \ast c$ for any invertible $c \in \mathbb{K}_n$. Then $H$ in proper $t$-Hessenberg form implies that the $t$-span $\left\{ \vec{G}_1, \vec{G}_2, \ldots, \vec{G}_j \right\} = \mathcal{K}_j(A, \vec{X})$ for $j = 1, \ldots, m$.

Next we want to verify that a tensor $QR$ algorithm implements nested subspace iteration for tensors, but several details remain. For instance, we require a more thorough understanding of roots of polynomials whose coefficients come from $\mathbb{K}_n$. Since $\mathbb{K}_n$ is not a field, the situation here is more complicated than for matrices. There is also the question of deflation: if $H$ is not in proper $t$-Hessenberg form, that is, if it has a non-zero but non-invertible subdiagonal tube, does it deflate? The answer to this question would shed light on not only a tensor bulge-chasing QR-type algorithm, but also on the issue of breakdown in a tensor Arnoldi method.
Fine Grained Sparse QR Factorization for Multicore Systems

Alfredo Buttari

Abstract

This work presents an efficient implementation of the sparse, multifrontal QR factorization for multicore systems based on fine-grained, dataflow parallelism.

The multifrontal method, introduced by Duff and Reid [6] as a method for the factorization of sparse, symmetric linear systems, can be adapted to the QR factorization of sparse matrices thanks to the equivalence of the R factor of a matrix $A$ and the Cholesky factor of the normal equation matrix $A^T A$ [1, 5]. At the heart of this method is the concept of elimination tree, a tree-structured graph describing the dependencies among computational tasks in the multifrontal factorization. Each node of the tree is associated to a dense matrix (commonly known as frontal matrix or simply front) and the whole factorization consists in a bottom-up traversal of the tree where, at each node two operations are performed:

- **assembly**: data produced by the processing of children nodes is assembled together with nonzero values from the original matrix to form the frontal matrix;

- **factorization**: the assembled frontal matrix is factorized. This step produces parts of the factors of the original matrix and a complement (commonly referred to as contribution block) which corresponds to the data that will be later assembled into the father node.

The multifrontal method can achieve very high efficiency on modern computing systems thanks to the fact that all the computations are arranged as operations on dense matrices; this reduces the use of indirect addressing and allows the use efficient Level-3 BLAS routines.

In a multifrontal factorization, parallelism is commonly exploited at two levels:

- **tree-level parallelism**: the multifrontal method lends very naturally to parallelization thanks to the fact that dependencies between computational tasks are captured by the elimination tree which identifies independent operations that can be, thus, performed in parallel;

- **node-level parallelism**: if the size of a frontal matrix is big enough, its factorization can be performed in parallel by multiple threads.

The classical approach to shared-memory parallelization for QR multifrontal solvers (see [1, 5]) is based on a complete separation of the two sources of concurrence described above. The node parallelism is delegated to multithreaded BLAS libraries and only the tree parallelism is handled at the level of the multifrontal factorization. This is commonly achieved by means of one, or more, task queues where a task corresponds to the assembly and factorization of a front. A new task is pushed into the queue as soon as it is ready to be executed, i.e., as soon as all the tasks associated to children nodes are completed. Threads keep polling the queue for tasks to perform until all the nodes of the tree have been processed. Although this approach works reasonably well for a limited number of cores, it suffers scalability problems mostly due to two factors. First, the separation of tree and node parallelism may cause load unbalance and may result in a suboptimal exploitation of the concurrency in the multifrontal method because of the highly variable granularity of computational tasks; second, large grained computational tasks introduce heavy synchronizations between threads which result in idle times in the factorization workflow.
This work [2] extends previous research on dense matrix factorizations [3, 4] and aims at overcoming the limitations of the classical approach discussed above by employing a different parallelization technique based on fine granularity partitioning of data combined with a data-flow model for the scheduling of tasks.

In order to handle both tree and node parallelism in the same framework, a 1D block-column partitioning of the frontal matrices is applied and three elementary operations defined:

1. **panel**: this operation amounts to computing the QR factorization of a block-column;

2. **update**: updating a block-column with respect to a panel corresponds to applying to the block-column the Householder reflections resulting from the panel reduction;

3. **assemble**: assembles a block-column into the father node (if it exists);

The multifrontal factorization of a sparse matrix can, thus, be defined as a sequence of tasks, each task corresponding to the execution of an elementary operation of the type described above on a block-column belonging to a front. The tasks are arranged in a Direct Acyclic Graph (DAG) of dependencies where edges define the order in which tasks have to be executed.

This DAG globally retains the structure of the elimination tree but expresses a higher degree of concurrency thanks to the fact that tasks are defined on a block-column basis instead of a front basis. This allows to handle both tree and node parallelism in a consistent way.

The execution of the tasks in the DAG is controlled by a data-flow model; a task is dynamically scheduled for execution as soon as all the input operands are available to it, i.e., when all the tasks on which it depends are completed. The scheduling of tasks can be guided by a set of rules that prioritize the execution of a task based on, for example, data locality in order to maximize the reuse of data into cache memories, or DAG nodes fan-out in order to generate more concurrency.

Experimental results show that, on a 24-cores AMD system, this approach delivers up to three times better scaling and performance than the recently released SuiteSparseQR [5] code based on the classical approach to parallelization.

**References**


Ritz Values of Nonsymmetric Matrices
and Convergence of the Restarted Arnoldi Method

Russell Carden, and Mark Embree

Abstract

The restarted Arnoldi method computes a few desired eigenvalues of a matrix by iteratively applying polynomial filters to refine eigenvalue approximations. Sorensen proposed using selected Ritz values – called “exact shifts” – as the roots of the filter polynomial. For the symmetric case, Sorensen proved that exact shifts ensure convergence as a consequence interlacing [5]. For nonsymmetric matrices the behavior of Ritz values is poorly understood, and hence no satisfactory general convergence theory exists. Though the method performs well in practice and can be found in Matlab’s `eigs` command and ARPACK, the algorithm can in theory fail when unwanted Ritz values coincide with eigenvalues [3]. As Ritz values are the basis for many iterative methods for eigenvalues and linear systems, an understanding of Ritz value behavior for nonsymmetric matrices has the potential to inform a broad range of analysis.

To better understand Ritz values, we study the numerical range of a matrix: \[ W(A) = \{ x^* A x : \| x \| = 1 \} \]. We demonstrate that Ritz values of nonsymmetric matrices cannot cluster throughout the entire numerical range. For example one can show that

\[ \sum_{j=1}^k \mu_j \leq \sum_{j=1}^k \Re \theta_j, \quad k = 1, \ldots, n - 1, \]

where \( \mu_j \) and \( \theta_j \) are the eigenvalues of the Hermitian part of \( A \) and the Ritz values from a subspace of dimension \( n - 1 \), ordered such that \( \Re \mu_j \leq \Re \mu_{j+1} \) and \( \Re \theta_j \leq \Re \theta_{j+1} \). That is, the real parts of the Ritz values from any subspace are weakly majorized by the eigenvalues of the Hermitian part of \( A \).

To study the interior structure of the numerical range we propose the inverse field of values problem (iFOV-\( k \)): given \( \{ z_1, z_2, \ldots, z_k \} \subset W(A) \), do there exist \( k \)-dimensional subspaces \( V \) such that \( \{ z_1, z_2, \ldots, z_k \} = \sigma(V^* A V) \) where \( \text{Ran}(V) = V, V^* V = I \)? For \( k = 1 \), the problem is always solvable and a solution can be constructed using Johnson’s algorithm for drawing the numerical range [6, 1]. However, for \( k > 2 \) the problem of determining if all the specified \( z_i \) can be the eigenvalues of some restriction of \( A \) is nontrivial. We look at this problem for both normal non-Hermitian matrices and Jordan blocks. For the normal case, there are known criteria for when the problem is solvable [4], we provide a simple geometric interpretation and find that Ritz values must obey constraints similar to interlacing [2]. For Jordan blocks we show that the Ritz values can indeed be localized, that is, the Ritz values cannot cluster throughout the numerical range.

At a more abstract level, we find that for a subspace to produce certain Ritz values, those Ritz values must lie in a certain algebraic variety. From the variety, one may derive sharp bounds on the location of all possible Ritz values. However, the most practical bounds generated thus far are those derived from the weak majorization given above. Using these bounds, we develop sufficient conditions that guarantees convergence of the restarted Arnoldi method with exact shifts for a matrix with a simple normal eigenvalue.
References


Gramian Based Model Reduction of Switched Dynamical Systems

Younès Chahlaoui

Abstract

We consider a switched linear dynamical system described by

\[
\begin{align*}
\delta x(t) &= A_\sigma x(t) + B_\sigma u(t), \\
y(t) &= C_\sigma x(t),
\end{align*}
\]

where \(x(t)\) is the state, \(u(t)\) is the controlled input, \(y(t)\) is the measured output, \(\sigma\) is the piecewise constant signal taking values from an index set \(M = \{1, \ldots, l\}\), and \(A_k, B_k\) and \(C_k, k \in M\) are matrices of appropriate dimensions. The switched system is a multi-model which is a special case of hybrid systems [3, 4].

This talk is about model reduction of switched systems which has received relatively little attention in the Numerical Linear Algebra community. We will present several new Gramian-based methods. These Gramians are matrix energy functions and they are, in theory, solutions of certain complicated Lyapunov equations. Here we propose to solve a set of simpler Lyapunov equations and to use linear combinations of these solutions to obtain the Gramians. We propose also a balanced truncation-like method with these two Gramians. We will also present another new algorithm based on Lyapunov stability analysis. We will show how to solve the underlying set of Linear Matrix Inequalities for two common solutions. These two solutions are used to come up with a balanced truncation-like method. With this approach we will preserve the stability for the reduced model.

We will suppose implicitly that each subsystem is stable.

The switching controlled by \(\sigma\) is introducing a form of uncertainty and the composite system comprises a certain number of subsystems where, at every time step, there is a certain probability that a particular subsystem will be switched on. Any operation on the composite system involves many numerical difficulties especially for stability or reachability analysis and formal verification [3, 4]. The use of numerical model reduction techniques has the potential to make feasible the computational investigation of a class of currently intractable systems. Gramian-based methods are one major class of the model reduction techniques [1, 5]. But it is a challenge, since almost all reduction methods cannot be directly applied to switched systems [2].

In general, a switching signal may depend on the time, its own past value, the state/output, and/or possibly an external signal as well

\[
\sigma(t+) = \sigma(t, \sigma(t), x(t)/y(t), z(t)) \quad \forall t
\]

where \(z(t)\) is an external signal produced by other devices, \(\sigma(t+)\) is the next switching signal. The switching is also making every operation on the system more complex as the number of involved subsystems is large. For example, a simple check of the stability, which for a single subsystem is equivalent to the eigenvalue problem of the dynamic matrix, is now a multiple eigenvalue problem of different matrix pencils. For instance, we have to check the eigenvalues of all linear combinations of each triplet \((A_i, A_j, A_k)\) and \((A_i, A_j^{-1}, A_k)\) for all \(i \neq j \neq k \in M\).

In the computer science community, which has for some time considered these systems, the switching signal has been classified according to different laws on time-driven or event-driven dynamics. In view of the applications we have in mind and for the purpose of using similar ideas that were developed for classic dynamical systems we will distinguish two cases for the switching signal: time-transmission switching and time independent switching.
For the time-transmission switching, the switching path is known a priori (i.e., we know a priori at what moment \( t_k \) the system is switching to which subsystem \( i_k \)). Given a switching sequence \( \{x_0, (t_0, i_0), (t_1, i_1), \ldots, (t_l, i_l)\} \), the state is given at any time \( t \in [t_k, t_{k+1}) \) by

\[
x(t) = \phi(t, t_0, x_0, u, \sigma) = \Phi(t, t_0, \sigma, x_0)x_0 + \sum_{j=1}^{k} \Phi(t, t_j, \sigma, x_0) \int_{t_{j-1}}^{t_j} e^{A_{i_{j-1}}(t_j-\tau)} B_{i_{j-1}} u(\tau) d\tau
\]

where the transition matrix

\[
\Phi(t_1, t_2, \sigma, x_0) = \psi(t_1, \sigma, x_0)(\psi(t_2, \sigma, x_0))^{-1} \quad \text{and} \quad \psi(t, \sigma, x_0) := e^{A_{i_k}(t-t_k)} \prod_{j=k}^{1} e^{A_{i_{j-1}}(t_j-t_{j-1})}.
\]

The key idea is to use the sub-Gramians, which are the Gramians (energy matrix functions) of each subsystem and solution of the Lyaponuv equations

\[
A_j\mathcal{P}_j + \mathcal{P}_jA_j^T + B_jB_j^T = 0, \quad A_j^T\mathcal{Q}_j + \mathcal{Q}_jA_j + C_j^TC_j = 0.
\]

The Gramians for the switched system are \( \mathcal{P} = \sum_{j=1}^{l} \alpha_j\mathcal{P}_j, \quad \mathcal{Q} = \sum_{j=1}^{l} \beta_j\mathcal{Q}_j, \) where \( \alpha_j \) and \( \beta_j \) define how much each subsystem is involved in the system. These parameters can be chosen following different strategies: for example \( \alpha_j = \beta_j = \frac{t_j-t_{j-1}}{t_l-t_0} \) or \( = \frac{1}{l} \) simply, if we know that each subsystem is visited at least once. Other strategies will be presented.

For the time-independent switching, the switching is based on the notion of guards. Once a guard is hit the switching occurs. To be able to reduce the system without destroying its nature (switching) we have to use properties of the system like the stability. The stability of the switched system is equivalent to the existence of two common solutions to the following systems of LMIs

\[
\mathcal{P} = \mathcal{P}^T \in \mathbb{R}^{N \times N}, \quad \mathcal{P} > 0, \quad A_\sigma\mathcal{P} + \mathcal{P}A_\sigma^T + B_\sigma B_\sigma^T < 0,
\]

\[
\mathcal{Q} = \mathcal{Q}^T \in \mathbb{R}^{N \times N}, \quad \mathcal{Q} > 0, \quad A_\sigma^T\mathcal{Q} + \mathcal{Q}A_\sigma + C_\sigma^T C_\sigma < 0.
\]

Until now, no efficient algorithm exists for solving these LMIs in the general case. Even for some special simple cases, solving numerically these LMIs is still a challenge once both \( N \) and \( l \) are large. We propose a new algorithm to compute efficiently \( \mathcal{P} \) and \( \mathcal{Q} \).

References


S. Chandrasekaran, M. Gu, H. Mhaskar, J. Moffitt and K. Raghuram

Abstract

It is well-known that classical Lagrange interpolation will produce polynomials that can diverge from the true function as the number of sample points increase, even if the true function is very smooth. We show that a simple modification of the interpolation equations will yield interpolating polynomials that converge for arbitrary point distributions even in higher-dimensional spaces. We call this approach the Minimum Sobolev Norm (MSN) technique.

With this approach in hand we can now re-visit classical approaches to the design of differentiation and quadrature rules, with greater freedom to construct high-order rules that are free to use scattered data points in high-dimensions.

We show one example of this by constructing higher-order finite-difference schemes on scattered data, and use this to construct an end-to-end 2D elliptic PDE solver. We compare both accuracy and efficiency of our code with conventional finite-difference and finite-element solvers. Our algorithm can achieve much higher-orders of accuracy and be significantly faster than existing codes.

The MSN approach to computing higher-order finite-difference weights involves the solution of diagonally weighted least-squares problems, with the weights being highly ill-conditioned. We will show how variants of the Hough-Vavasis algorithm can be used to solve these problems to almost full-accuracy. In particular we will present a poor man’s version of the Hough-Vavasis algorithm, that can be implemented using an SVD code in place of an RRQR code.

It is well-known that finite-difference and finite-element methods produce linear systems of equations whose condition number can grow quite rapidly with problem size. This means that with decreasing grid size the numerical solution can actually start to diverge. While conventionally the community has worried about the effect of these large condition numbers on the convergence rate of iterative methods, here we will concentrate on its effects on accuracy, since we are using a direct sparse solver in our code. We will show how these systems can be solved so as to mitigate the loss of accuracy for small grid sizes.
LLL Reduction and Integer Least Squares Problems

Xiao-Wen Chang

Abstract

We summarize our recent work on LLL reduction and integer least squares (ILS) problems. Given a matrix \( A \in \mathbb{R}^{m \times n} \) with full column rank and a vector \( b \in \mathbb{R}^m \), the ordinary integer least squares (ILS) problem has the form:

\[
\min_{x \in \mathbb{Z}^n} \| b - Ax \|_2^2.
\]  

(1)

Sometimes \( x \) may be subject to some constraints and \( A \) may be underdetermined (see, e.g., [4], [5], [7], [12]). Since the ILS problem is to find a point in the lattice \( \mathcal{L}(A) = \{ Ax : x \in \mathbb{Z}^n \} \) which is closest to \( b \), the ILS problem is also referred to as the closest point problem in lattice theory. Integer least squares (ILS) problems arise from many applications such as communications, cryptoanalysis, global positioning systems (GPS), lattice design, radar imaging, etc; see, e.g., [1], [12], [13], and [14]. Unlike real LS problems, the general ILS problems are NP-hard. A typical approach to solving an ILS problem consists of two phases: reduction and search; see, e.g., [1]. The reduction process (which is often referred to as the decorrelation process in the GPS literature) transforms the original problem to a new problem to make the search process easier and more efficient and the search process enumerates the integer points over a specific region and finds the optimal solution. Currently the search strategy proposed in [13] is widely used.

The reduction process for the ordinary ILS problem (1) can be described by what we called the QRZ factorization (see [4]):

\[
A = Q \begin{bmatrix} R \\ 0 \end{bmatrix} Z,
\]

where \( Q \in \mathbb{R}^{m \times m} \) is orthogonal, \( R \) is upper triangular, and \( Z \in \mathbb{Z}^{n \times n} \) is unimodular (i.e., \( \det(Z) = \pm 1 \)). In [2] we have given some new insights about what properties of \( R \) the reduction process should try to achieve to make the search process faster. We used some examples to illustrate two misconceptions in the literature: \( R \) should be as well-conditioned as possible and \( R \) should be as diagonal as possible. In the QRZ factorization, if \( R \) satisfies the following properties

\[
\begin{align*}
 r_{kk}^2 &\leq \delta (r_{k,k+1}^2 + r_{k+1,k+1}^2), \\
 |r_{kj}| &\leq |r_{kk}|/2, \quad j = k + 1 : n, \quad k = 1 : n - 1
\end{align*}
\]

(2)

where \( \delta \) is a constant and \( 1 \leq \delta < 2 \), then \( AZ^{-1} \) is said to be LLL reduced. The celebrated LLL reduction [11] has been widely used in solving (1) (see, e.g., [1], [8], [9] and [10]). We showed that for solving the ILS problem (1), the second property in (2) for \( j = k + 2 : n \) is not needed, because the cost of the search process (suppose we use the search process given in [13]) does not depend on this property. This has led us to design a new faster partial LLL reduction algorithm. In designing the new algorithm, numerical stability was taken into account.

In some applications one wants to determine if a given integer matrix is LLL reduced. We can try to compute the QR factorization of the given matrix by floating point arithmetic and check if the computed \( R \) factor satisfies the properties in (2). But then we need to investigate the accuracy of the computed \( R \) factor. In [6] we gave rigorous error bounds on the \( R \) factor. The results should be very useful to devise LLL-type algorithms relying on floating point approximations.

Recently, when we studied synchronous rate and frequency adjustment for utilization regulation in distributed real-time embedded systems [3], we encountered the following mixed-discrete bilinear
LS problem:
\[
\min_{X,y} \|b - XAy\|_2,
\]
where \(b \in \mathbb{R}^m\), \(X = \text{diag}(x_1, \ldots, x_m) \in \mathbb{R}^{m \times m}\) with \(l \leq x_i \leq u\), \(A \in \mathbb{R}^{m \times n}\), \(y \in \mathbb{R}^n\) with \(y_i \in \{c_{i1}, c_{i2}, \ldots, c_{il}\}\). In [3] we proposed an alternating iterative method to solve this problem.

References


Error Analysis for Nonlinear Model Reduction using POD-DEIM Technique

Saifon Chaturantabut, and Danny C. Sorensen

Abstract
Model order reduction (MOR) has been extensively used for efficient simulation of dynamical systems in various applications. To assess the quality of a MOR technique, it is important to analyze the accuracy and stability of the resulting reduced system. This work focuses on the error analysis of solutions from the reduced systems constructed by Proper Orthogonal Decomposition (POD) together with Discrete Empirical Interpolation Method (DEIM) for nonlinear dynamical systems.

The POD-DEIM technique can be viewed as a variation of the standard POD-Galerkin approach with an improvement on efficiency in computing nonlinear terms. A reduced system constructed from the POD-Galerkin approach often has a small number of unknowns, but still has complexity required for evaluating the projected nonlinear term proportional to the dimension of the original system. DEIM can be used for approximating the nonlinear term to significantly reduce this complexity further. The success of POD-DEIM approach is recently illustrated through the applications such as neural modeling [2] and two-phase miscible flow in porous media [3].

The properties such as accuracy, convergence and stability of the reduced system from the POD-Galerkin approach have been studied in a number of works. However, it appears that the analysis of DEIM can only be found in [1] where an error bound for the DEIM approximation of a nonlinear vector-valued function was shown to be nearly as accurate as the optimal orthogonal projection in POD approximation. This error bound will be further used in this work to obtain the overall error involved in solving a nonlinear dynamical system.

In this work, we investigate the accuracy of POD-DEIM reduced system for nonlinear ordinary differential equations. The asymptotic error analysis will be first considered in the continuous setting where the overall accuracy of the reduced system is only contributed from applying the POD-DEIM technique without other effects including the choice of time integration method. Then, a framework for error analysis in the discrete setting for the implicit Euler time integration scheme will be presented. The proposed error bounds in both continuous and discrete settings are derived through an application of generalized logarithmic norms for unbounded nonlinear operators. The conditions under which the stability of the original system is preserved and the reduction error is uniformly bounded will be discussed. The resulting error bounds in 2-norm are shown to be proportional to the sum of the singular values corresponding to neglected POD basis vectors both in Galerkin projection of the reduced system and in DEIM approximation of the nonlinear term.

References


Windowed Spectral Regularization of Inverse Problems

Julianne Chung, Glenn Easley, and Dianne O’Leary

Abstract

Ill-posed inverse problems arise in a variety of scientific and engineering applications. The discretized problem can often be written as a linear system

\[ b = Ax_{\text{true}} + \varepsilon, \]  

where \( b \in \mathbb{R}^m \) is the vector of observed data, \( x_{\text{true}} \in \mathbb{R}^n \) is the desired true solution, matrix \( A \in \mathbb{R}^{m \times n}, m \geq n, \) is known, and \( \varepsilon \in \mathbb{R}^m \) represents noise in the data. The goal is to compute an approximation of \( x_{\text{true}} \), given \( b \) and \( A \).

Most inverse problems are ill-posed, meaning that small perturbations in the data may result in large errors in the solution [2, 7]. Regularization via spectral filtering can be used to filter out “hazardous” components (components that highly perturb the estimate) of the solution in order to compute stable solutions. Standard filtering methods, such as the truncated singular value decomposition and Tikhonov regularization, work in the frequency domain of the discretized kernel, defined by the singular value decomposition (SVD) of the matrix \( A \). However, a key limitation is the use of a single regularization parameter to define the shape of the filter and ultimately to determine the quality of the solution [5].

In this presentation, we present a novel windowed approach for spectral regularization, where windows break the problem into subproblems with narrower bands of singular values, and each subproblem is regularized individually. Let \( A = U \Sigma V^T \) be the SVD of \( A \), and let \( w^{(j)} \in \mathbb{R}^n, j = 1, 2, \ldots, p, \) be nonnegative frequency weights such that \( \sum_{j=1}^{p} w^{(j)}_i = 1 \). Let \( \lambda^{(j)} \) be the regularization parameter corresponding to window \( W^{(j)} = \text{diag}(w^{(j)}) \). Then the windowed reconstruction can be computed as

\[ x_{\text{win}} = \sum_{j=1}^{p} V \left( \Sigma^T \Sigma + \lambda^{(j)} I \right)^{-1} W^{(j)} \Sigma^T U^T b. \]

The novelty of our approach is in applying varying amounts of regularization to the observed data at different frequency scales. The advantages of our approach include separation of local features in the data and better conditioning within each window or band of frequencies. In addition, a perturbation analysis shows that we can get improved error bounds and improved conditioning in each band of the windowed approach compared to standard methods. Furthermore, this approach can be robust to high noise levels or colored noise. The cost of these improvements is the need to select multiple regularization parameters. Some research has been done on selecting multiple regularization parameters [1, 4, 5, 6, 8]; however, none of these previous approaches considers decompositions in the frequency domain of the operator.

We consider a variety of overlapping and non-overlapping approaches for constructing the spectral windows, \( W^{(j)} \). We extend standard parameter-choice methods, such as the discrepancy principle and generalized cross-validation [3], to a windowed regularization framework. Numerical results demonstrate the effectiveness of our algorithms on deblurring images and on the backward heat equation, and investigations are made into how the number of windows and noise level affect solutions.
References


Abstract

The Eckart-Young theorem characterizes the minimum norm properties of the SVD of a given matrix, $A$. It says that the nearest rank-$k$ matrix is provided by the corresponding truncated SVD. This feature forms the basis of several matrix approximation methods. The Ky Fan’s maximum principle is another example of a well-known useful result. Given a symmetric matrix $S$, it considers the problem of finding a symmetric Rayleigh quotient matrix of $S$ that has the largest trace. These two fundamental observations are always presented separately, as independent and unrelated results. However, as this talk shows, the two theorems are closely related.

The Orthogonal Quotients Equality that we prove turns the Eckhart-Young theorem from a minimum norm problem into an equivalent maximum norm problem. The new formulation considers the problem of finding an orthogonal quotient matrix of $A$ that has the largest Frobenius norm. The symmetric version of the Orthogonal Quotient Equality turns Ky Fan’s maximum principle into a minimum trace problem. A comparison of the extended formulations reveals a surprising similarity between Eckart-Young theorem and Ky Fan’s maximum principle. The similarity suggests that the two theorems are reflecting “two sides of the same coin”. That is, both theorems are special cases of a more general extremum principle.

The second part of the talk is aimed at proving this conjecture. For this purpose we derive a new extremum principle. One that considers the problem of maximizing (or minimizing) unitarily invariant norms of orthogonal quotient matrices. In this case the objective function is often expressed as sum of powers of singular values. (The singular values of orthogonal quotients matrices of $A$.) The proofs are based on “rectangular” versions of Cauchy interlace theorem and Poincare’ separation theorem. The extended theorems have several useful conclusions. In particular, it is shown that both the Eckart-Young theorem and Ky Fan’s maximum principle are special cases of the extended extremum principle.

References


Avoiding Communication in Numerical Linear Algebra

James Demmel

Abstract

Algorithms have two kinds of costs: arithmetic and communication, by which we mean moving data either between levels of a memory hierarchy (in the sequential case) or between processors over a network (in the parallel case). Communication costs can already exceed arithmetic costs by orders of magnitude, and the gap is growing exponentially over time, so our goal is to design linear algebra algorithms that minimize communication. First, we show how to extend known communication lower bounds for $O(n^3)$ dense matrix multiplication to all direct linear algebra, i.e. for solving linear systems, least squares problems, eigenproblems and the SVD, for dense or sparse matrices, and for sequential or parallel machines. Second, we overview new algorithms that attain these lower bounds, almost none of which are in standard libraries; some of these algorithms are variations on conventional ones, and some are quite different. We will present both performance models and measured timings for some implementation showing that large speedups over conventional algorithms are possible. Third, we describe extensions of these results to Strassen-like algorithms. Fourth, we show how to minimize communication in Krylov-subspace methods for solving sparse linear system and eigenproblems, and again describe new algorithms with significant speedups.

The details of many of these new algorithms, as well as lower bounds for Strassen-like algorithms, will be presented by collaborators who are also submitting abstracts to this symposium, including Grey Ballard, Erin Carson, Ioana Dumitriu, Laura Grigori, Magnus Gustafsson, Mark Hoemmen, Nick Knight, Julien Langou, Oded Schwartz, and Edgar Solomonik. This talk is meant to be an overview of this entire body of work, with more details about certain topics, including the lower bounds on $O(n^3)$-like (non-Strassen-like) algorithms.
Convergence Bounds for Approximate Invariant Subspace Recycling for Sequences of Linear Systems

Eric de Sturler

Abstract

Recycling selected subspaces in a sequence of linear systems to reduce the overall iteration count has proven very effective in a range of applications, such as computational mechanics, acoustics, optimal design, uncertainty quantification, diffuse optical and electrical impedance tomography, model reduction, and solving generalized eigenvalue problems. Several types of subspaces, such as invariant subspaces, solution spaces, spaces based on the underlying problem, and combinations of these have been used successfully for recycling. Approximate invariant subspaces have proved successful for many problems, and empirical observation suggests that very modest accuracy is sufficient to obtain convergence close to that for exact deflation. For our purposes, efficiently solving a sequence of changing linear systems, this property is of paramount importance. Hence, we focus on the analysis of recycling an approximate invariant subspace. Although our interest is in solving a sequence of changing linear systems, our results are generally applicable to Krylov methods where an approximate invariant subspace is projected out in each iteration.

Quite a few authors have provided convergence analyses for Krylov subspace methods iterating orthogonal to an (approximate) invariant subspace, for example [Saad, SIMAX’97], [Vuik et al, TR 10-14], [Nabben and Vuik, NLAA’08], and [Bai and Meerbergen, SIMAX’10]. Simoncini and Szyld [SIREV’05] provide a convergence analysis for standard Krylov methods that provides insight for this case as well, and the analysis of seed methods by Chan and Ng [SISC’99] also provides some insight for this case. However, for optimal Krylov methods projecting out (recycling) an approximate invariant subspace, a concise and clear quantitative statement that links the convergence bounds to the accuracy of this approximation and explains the empirical observation that only very modest accuracy is required to obtain convergence as for exact deflation appears still to be missing. We will provide such a result for both Hermitian and non-Hermitian matrices. We note that the nice results by Saad (above) are related to ours, but they require or suggest the need for much greater accuracy of the approximate invariant subspace.

In this abstract, we briefly discuss the result for the Hermitian, possibly indefinite, case (without proof) and point to some similarities in the result for non-Hermitian systems. We will discuss both proofs and consequences in the presentation. As always, the results are somewhat weaker for the non-Hermitian case than for the Hermitian case, but they confirm that, in this case too, modest approximations, which are easily obtained, are sufficient except for pathological cases.

Consider solving the Hermitian linear system $Ax = b$, and a matrix, $U$, whose $k$ columns span the recycling space. For simplicity, assume $U$ has already been computed such that $AU = C$ has orthonormal columns. A recycling MINRES would be implemented as follows. Take $x_0 = UC^*b$, $r_0 = (I - CC^*)b$, set $v_1 = r_0/\|r_0\|$, and carry out the Lanczos iteration $(I - CC^*)AV_m = V_{m+1}\tilde{T}_m$, where $\tilde{T}_m$ is an $(m+1) \times m$ tridiagonal matrix. Approximations from this iteration are computed (more or less) in the usual way [Wang and de Sturler, IJNME’07]. Note, that since $V_{m+1} \perp C$ by construction, this is equivalent (but cheaper) to iterating with $(I - CC^*)A(I - CC^*)$. We can bound the convergence by considering the eigenvalues of this matrix.

We consider the typical case that we have a few ‘problematic eigenvalues’ that are small in absolute value and/or make the matrix indefinite. Let $A$ have the eigendecomposition $A = [Q Y] \text{diag}(\Lambda_Q, \Lambda_Y)[Q Y]^*$, where $\Lambda_Q$ is a set containing the $\ell \leq k$ ‘problematic eigenvalues’ we
would like to remove, \( \Lambda_Y \) is the set of remaining eigenvalues, \( Q \) has \( \ell \leq k \) orthonormal eigenvectors of \( A \) corresponding to \( \Lambda_Q \) as its columns, and \( Y \) contains the remaining orthonormal eigenvectors as its columns. Now, let \( \text{Range}(U) \) be an approximate invariant subspace in the following sense, \( \| (I - CC^*)Q \|_2 = \delta < 1 \), that is, the invariant subspace \( \text{Range}(Q) \) is approximately contained in the space we project out of the Krylov iteration (where \( \delta \) is small but not necessarily very small).

Now let \([C W]\) be a unitary matrix. Then our recycling MINRES above iterates over the space \( \text{Range}(W) = \text{Range}(I - CC^*) \), and its convergence is governed by the Rayleigh quotients (or field of values), \( z^*WW^*AWW^*z \) for unit \( z \in \text{Range}(W) \). We have

\[
\delta^2 \min\{0, \Lambda_Q\} + (1 - \delta^2) \min\{\Lambda_Y\} \leq z^*WW^*AWW^*z \leq \max\{\Lambda_Y\},
\]

and we obtain convergence bounds from these eigenvalue bounds in the usual way. For CG with recycling the same result is obtained (although the recycling must be implemented slightly differently). The eigenvalue bound shows that, for small \( \delta \), the field of values is a slight perturbation of the range of the remaining eigenvalues, and for \( \delta \to 0 \) the bound gives the obvious result that Rayleigh quotient ranges between \( \min(\Lambda_Y) \) and \( \max(\Lambda_Y) \). Particularly important for our purposes is that even for modestly small \( \delta \), say \( \delta = O(0.1) \) or \( \delta = O(0.01) \), the bound indicates a significant improvement in the upper bound on the convergence rate. In fact, for such \( \delta \), the convergence bound is already essentially the same as for exact deflation, unless \( \min(\Lambda_Q) \) is negative with large magnitude. Note that such modestly small \( \delta \) reflects the fact that very modest accuracy of the invariant subspace is sufficient, confirming empirical observations. Such accuracy is easy to get in practical situations, even if the matrix is changing [Parks et al, SISC’06]. Furthermore, in the HPD case, the left eigenvalue bound is just slightly smaller than the minimum of \( \Lambda_Y \), even for, say, \( \delta = 0.1 \). In terms of approximating an invariant subspace that is a very modest accuracy. In the case of negative eigenvalues, \( \delta \) may need to be somewhat smaller for a good convergence bound (depending on the magnitude of the negative eigenvalues and \( \min(\Lambda_Y) \)).

Although we apply our result here to a specific choice of invariant subspaces, our approach is general. We also note that, for practical purposes, the number of ‘problematic eigenvalues’ should not be too large, otherwise the recycling becomes too expensive.

For the non-Hermitian case, we obtain similar results for bounds based on the field of values and on the pseudo-spectrum. These bounds, too, are small perturbations of bounds based on the field of values or pseudo-spectrum for the complementary invariant subspace, including effects of nonnormality. In this case, the perturbation is based on two parameters, the same \( \delta \) as defined above and the smallest canonical angle, \( \vartheta \), between the spaces \( \text{Range}(Q) \) and \( \text{Range}(Y) \). One nice result is that the bound shows how \( \delta \) needs to decrease as a function of \( \vartheta \) to keep the perturbation small. The other important result is that the bound shows that modest accuracy of the invariant subspace, say \( \delta = O(0.01) \) is sufficient unless \( \vartheta \) is very small. It is important to note that even for highly nonnormal matrices this bound can be quite good as long as the canonical angles between the actual spaces \( \text{Range}(Q) \) and \( \text{Range}(Y) \) are not very small (see Simoncini and Szyld, above, for similar observations).

Finally, we note that these bounds have wider applicability. For example, they can be used as alternatives for convergence analyses of methods like CG and GMRES (without recycling) in a similar way as the analyses by van der Sluis and van der Vorst [Num. Math.’86], van der Vorst and Vuyk [JCAM’93], and Simoncini and Szyld [SIREV’05]. It will be interesting to see how these results compare and how the various approximations are linked.
Congruence Orbits of Matrices and Palindromic Pencils, and the Solution of the Equation $XA + AX^T = 0$

Fernando de Terán, and Froilán M. Dopico

Abstract

Both the computation of the canonical form for similarity of matrices (the Jordan canonical form) and of the canonical form for strict equivalence of matrix pencils (the Kronecker canonical form) are ill-posed problems due to the presence of arbitrarily close different structures to a given one. The theory of orbits provides a geometric interpretation and, moreover, introduces a theoretical framework in the problem of the computation of these canonical forms. Apart from its theoretical relevance, this framework may be useful in the development of algorithms and the explanation of their eventual failures.

Thanks to the work of a series of authors (we just cite some of the most relevant ones [1, 8, 3, 4, 5] but the complete list is larger), nowadays we know quite in detail the orbit space for similarity of matrices and the orbit space for strict equivalence of matrix pencils. In particular, we know not only which are all the nearby structures to a given one, but also the inclusion relationship (hierarchy) between the orbit closures.

By contrast, a parallel theory has not been addressed so far for the case of congruence orbits of matrices. In this talk, we present a first step in this direction. More precisely, we compute the dimension of the congruence orbits and we determine the generic canonical structure by congruence. For this purpose, we solve the equation $XA + AX^T = 0$, after showing that the codimension of the solution space of this equation coincides with the dimension of the congruence orbit of $A$.

Our interest in addressing the theory of congruence orbits is motivated by its close relationship with the theory of strict equivalence orbits of palindromic pencils. In the past few years, this kind of pencils has drawn the attention of researchers in the numerical linear algebra community because of their appearance in several applied problems (see, for instance, [7, 6]). We will show that the theory of congruence orbits of matrices and the theory of congruence orbits of palindromic pencils are equivalent theories. This will allow us to determine the dimension of congruence orbits of palindromic pencils. Moreover, we will show that the strict equivalence invariants of palindromic pencils coincide with the congruence invariants and, as a consequence, we will determine the generic Kronecker canonical form of these pencils.

The content of this talk is part of the work [2].

References


Social Network Analysis: Fast and Memory-Efficient Low-Rank Approximation of Massive Graphs

I. Dhillon, B. Savas and Y. Zhang

Abstract

Massive graphs arise routinely nowadays in social network analysis. For example, Facebook has over half-a-billion users, and the friendship relationships between its users can be captured by a massive graph. It is crucial to be able to analyze such massive graphs efficiently for various purposes. As an example, consider the Katz measure \[3\] which can be used for friend recommendations:

\[
\sum_{i=1}^{\infty} \beta^i A^i = (I - \beta A)^{-1} - I, \quad \beta < 1/\|A\|_2.
\]

Here, the matrix \(A\) is the adjacency matrix that captures friendship relationships, so the powers \(A^i\) capture information about paths of length \(i\) in the graph, while \(\beta^i\) is an exponentially decreasing damping factor. The Katz measure is often computed to predict future friend recommendations [5]. However, if \(A\) has dimension over 500,000 then computing the Katz measure can be prohibitively expensive. Many other problems in social network analysis rely on computations with the social graph.

Low-rank approximations are an obvious way to approximate massive graphs and perform tasks such as link prediction, collaborative filtering and semi-supervised learning on these graphs. However, traditional Lanczos-based methods to compute the truncated SVD can be both time and memory inefficient. The latter issue is especially troublesome as the matrix has huge dimension and the singular vectors are dense. In this talk, I will introduce Clustered Graph Embeddings, which provide an alternate, low-rank approximation scheme, which is (i) memory-efficient, and (ii) faster than computing a truncated SVD of the entire graph. Our proposed scheme relies on a fast clustering of the graph, which can be provided by fast multilevel methods such as Graclus [1] or Metis [4], followed by computation of separate low-rank approximations of these clusters (yielding sparse basis vectors), and then “gluing” together the approximations of the clusters to form an overall approximation of the entire graph. Within the clusters, low-rank approximations can be computed either by computing smaller truncated SVDs or employing stochastic low-rank approximations [2]. Theoretically, we generalize the bounds of [2] to our setting, which justifies the use of our clustering methods. Experimentally, we show that our method greatly outperforms the truncated SVD and stochastic low-rank approximations of the entire graph in terms of approximation error. We also show that the proposed method outperforms existing methods for friendship recommendation in massive social networks [6].

Our method is eminently suitable for parallel computation, especially on modern multi-core architectures, as it breaks down a large problem into smaller pieces. If time permits, I will discuss (i) parallel implementation details, (ii) development of a hierarchical version of the method, (iii) applications to settings other than social networks, namely, gene and disease/phenotype networks (collaboration with biologist Edward Marcotte; see www.nytimes.com/2010/04/27/science/27gene.html for more on his research), and call-graph networks on real data made available to UT Austin by the leading cell phone provider in Korea, South Korea Telecommunications (SKT).
References


Square Smoothing Operators Imposing Accurate Boundary Conditions

Marco Donatelli and Lothar Reichel

Abstract

We are concerned with large-scale discrete ill-posed problems, i.e., with the computation of an approximate solution of linear systems of the form

$$Ax = b, \quad A \in \mathbb{R}^{n \times n}, \quad x, b \in \mathbb{R}^n,$$

where $A$ is severely ill-conditioned and may be singular. These problems typically arise from discretizations of Fredholm integral equation of the First kind, e.g., in computerized tomography, geophysics or image restoration. Due to the ill-conditioning of $A$ and the unavoidable errors in the right-hand side (coming from data), it is necessary to resort to regularization.

One of the most popular regularization methods, known as Tikhonov regularization, replaces the linear system (1) by the minimization problem

$$\min_{x \in \mathbb{R}^n} \{\|Ax - b\|_2 + \mu\|Lx\|_2\}$$

where the matrix $L \in \mathbb{R}^{k \times n}, k \leq n$, is referred to as the regularization operator and the scalar $\mu \geq 0$ as the regularization parameter. The regularization operator $L$ should be chosen such that known features of the desired solution can be represented by vectors in the null space of $L$. When the solution is smooth, common choices of regularization operators in one space-dimension are scaled finite difference approximations of a derivative. Large-scale problems typically are transformed to standard form before solution [4]. This transformation requires the $A$-weighted pseudoinverse of $L$, which reduces to its inverse if $L$ is a square nonsingular matrix. Recent results show that Tikhonov regularization based on the range restricted Arnoldi process can be competitive with Tikhonov regularization based on Lanczos bidiagonalization and it does not require the evaluation of matrix-vector products with $A^T$. The disadvantage is that it requires the regularization operator $L$ to be represented by a square matrix.

For large-scale problems we need iterative methods to compute regularized solutions. Many iterative methods exhibit a “self-regularity property” in the sense that early termination of the iterative process has a regularizing effect. The approximations seem to converge in the initial iterations, before they become worse and finally diverge. In order to accelerate the convergence without spoiling the restoration, right smoothing preconditioners like the $A$-weighted pseudoinverse can be used (see [4]). Square smoothing regularization operators can be effective right preconditioners for some Krylov subspace methods like the GMRES method and the closely related Range Restricted GMRES (RRGMRES) [1].

In this talk, we discuss how to define square smoothing operators imposing appropriate boundary conditions to finite difference approximations of a derivative. We describe in detail the spectral decomposition of the matrices obtained imposing boundary conditions to the finite difference discretization of the second derivative as in [6]. Classical choices are Dirichlet and Neumann boundary conditions (see [6]). A natural discrete counterpart of Neumann boundary conditions is a symmetry around the midpoint which leads to a matrix diagonalized by the discrete cosine transforms (DCT). The second difference matrix is invertible in the case of Dirichlet boundary conditions, while it is singular with a null space that contains the constant functions in the case of Neumann boundary conditions.
Dirichlet and Neumann boundary conditions could be not appropriate if the solution has a different behavior at the boundary. Other discrete boundary conditions are the so-called “antireflective” boundary conditions, originally introduced for deblurring problems, which perform an antisymmetric extension of the solution that preserves linear functions. Such geometric property is especially evident from the spectral decomposition of the second difference matrix which has a null space containing constant and linear functions. The matrix-vector product with the matrix of the eigenvectors can be computed by the discrete sine transform (DST) up to lower order operations. The same holds for its inverse. Therefore, antireflective boundary conditions combine fast computations by DST with accurate restorations since they preserve constant and linear functions (see [3]).

When the solution does not have linear components, but for instance quadratic or exponential components, the eigenvector basis can be constructed combining a fast trigonometric transform with other solution components. This is the basic idea of the high order boundary conditions defined in [2] and it agrees with the last sentence in [6]: “We hope that the eigenvector approach will suggest more new transforms, and that one of them will be fast and visually attractive.”. Indeed, we note that such boundary conditions do not have any physical meaning but they are motivated only from the spectral decomposition of the solution.

The square smoothing operators obtained imposing boundary conditions share the properties that are suitable for fast computations (by fast trigonometric transforms) and their spectral decomposition is explicitly known. The latter allows to manipulate some eigenvalues and/or eigenvectors such that the modified operator has the desired behavior. In particular, we handle square smoothing regularization operators with appropriate boundary conditions investigating and proposing ad-hoc modifications in the following two directions:

1. Invertible operators to be used as right preconditioner with RRGMRES, LSQR, etc.
2. Operators with a good approximation of a prescribed null space and such that the matrix-vector product with its pseudoinverse can be computed by fast trigonometric transforms.

Numerical results and comparisons with other approaches to construct square smoothing operators proposed in [1, 5] confirm the effectiveness of our proposal.

References


Structured Perturbation Theory of Diagonally Dominant Matrices and Numerical Applications

Froilán M. Dopico

Abstract

Diagonally dominant matrices are among the most important classes of matrices arising in applications. For instance, the matrices that appear in many finite element methods are diagonally dominant and the same happens in finite difference methods for partial differential equations. Diagonally dominant matrices enjoy excellent theoretical and numerical properties that are explained in detail in standard references on Numerical Linear Algebra. However, if the condition number of a diagonally dominant matrix $A$ is very large, then conventional algorithms compute its singular value decomposition (SVD) and solve linear systems $Ax = b$ with very large relative errors.

In this talk, given a diagonally dominant matrix $A$, we consider perturbations that preserve this structure and present structured perturbation bounds for

- the LU factorization of $A$,
- the inverse of $A$,
- the solution of $Ax = b$, and
- the SVD of $A$.

To obtain these bounds, we make use of a proper parametrization of $A$ previously introduced in [3]. The new bounds do not depend on the traditional condition number of the matrix and are much better than standard perturbation bounds for arbitrary perturbations.

These perturbation results allow us to prove rigorously that certain algorithms compute the solution of linear systems $Ax = b$ accurately (for most vectors $b$) [2] and compute the SVD of $A$ with high relative accuracy [3], for any diagonally dominant matrix $A$ in $O(n^3)$ flops. Part of the results we plan to present can be found in [1], while other are still under development.

References


Numerical Implementation of the Iterative Rational Krylov Algorithm for Optimal $\mathcal{H}_2$ Model Reduction

Christopher Beattie, Zlatko Drmač, and Serkan Gugercin

Abstract

The Iterative Rational Krylov (IRKA) algorithm for model order reduction (Gugercin, Antoulas, Beattie 2008.) has recently attracted attention because of its effectiveness in real world applications, as well as because of its mathematical elegance. Given a LTI system (SISO, for simplicity) $(A, b, c)$, IRKA determines a reduced order system $(A_r, b_r, c_r)$ that satisfies certain $\mathcal{H}_2$ optimality conditions. The core iterations in IRKA are the fixed point iterations

\begin{equation}
\sigma^{(k+1)} = \phi(\sigma^{(k)}),
\end{equation}

where, for a selection of (generally complex) shifts $\sigma = (\sigma_1, \ldots, \sigma_r)$,

\begin{equation}
\phi(\sigma) = -\text{eig}(A_r(\sigma)) \equiv -\lambda(\sigma),
\end{equation}

\begin{equation}
A_r(\sigma) = (W(\sigma)^T V(\sigma))^{-1} W(\sigma)^T A V(\sigma)
\end{equation}

and $V(\sigma) = ((\sigma_1 I - A)^{-1} b, \ldots, (\sigma_r I - A)^{-1} b), W(\sigma) = ((\sigma_1 I - A^T)^{-1} c, \ldots, (\sigma_r I - A^T)^{-1} c)$.

Our goal is efficient and numerically reliable mathematical software that implements the IRKA algorithm. The first step is, necessarily, a theoretical study of the algorithm. We analyze the convergence of fixed point iterations, in particular the morphology of the mapping $\sigma^{(k+1)} = \phi(\sigma^{(k)})$ (fixed points, periodic points and their classification). Other theoretical issues include perturbation theory to analyze stability of the shifts, revealing relevant condition numbers, Cauchy–like structure of certain key matrices, connection of the fixed point iterations (1, 2) and pole placement, proper stopping criterion that translates into a backward stability relation, etc.

Besides rich theory, IRKA also offers many numerical challenges. How to implement the algorithm efficiently using direct solvers for $V(\sigma), W(\sigma)$? How to adapt iterative solvers in an inner loop that communicates with the outer fixed point iterations loop? How to implement $\phi(\sigma)$ efficiently? When to stop? All these and many other questions are analyzed during software development. In our presentation, we will give some detailed answers and illustrate the performances of the software.

References:


2. C. Beattie, Z. Drmač and S. Gugercin, An implementation of the IRKA algorithm for optimal $\mathcal{H}_2$ model reduction, in preparation.
Optimal Rational Krylov Subspaces for Large-Scale Dynamical Systems

Vladimir Druskin, Leonid Knizhnerman, Valeria Simoncini and Mikhail Zaslavsky

Abstract

Ruhe’s Rational Krylov Subspace is recognized as a powerful tool within Model Order Reduction techniques for linear dynamical systems. However, its success has been hindered by the lack of procedures, which would generate the sequence of shifts used to build the space with good approximation properties.

We begin with the first order passive problem

\[ Au + u_t = 0, \quad u|_{t=0} = b \]

for \( 0 \leq t < \infty \), where \( u(t), \varphi \in \mathbb{R}^N \) and \( A \in \mathbb{R}^{N\times N} \). We will solve this problem by projecting it onto the Rational Krylov Subspace (RKS)

\[ U = \text{span} \left\{ (A + s_1 I)^{-1} b, (A + s_2 I)^{-1} (A + s_1 I)^{-1} b, \ldots, \prod_{j=1}^{n} (A + s_j I)^{-1} b \right\}, \]

for some shifts \( -s_i \notin W(A) \) (the numerical range of \( A \)). We assume that the RKS is computed using iterative methods for which there are no computational advantages in solving multiple linear systems with the same shifts. Such problems often appear from large scale discretizations of PDEs in \( \mathbb{R}^3 \). Therefore, we do not require that the shifts coincide.

We first assume that \( W(A) \) is known and design a-priori algorithms of optimal shift generation. We consider this problem in the frequency domain and reduce it to the third Zolotaryov problem in complex plane

\[ \min_{s_1, \ldots, s_n} \max_{z \in \partial W(A)} \left| \frac{h_n(z)}{h_n(-z)} \right|, \]

where \( h_n(s) = \prod_{i=1}^{n} (s - s_i) \). A related approach using Blaschke product technique was suggested in [3]. The Zolotaryov problem yields an asymptotically optimal solution with \( s_i \in \partial W(A) \). We give explicit formulas for the error estimate and optimal shifts for symmetric \( A \), i.e., when \( W(A) \) is a real interval [4]. We also construct an infinite sequence of shifts yielding a nested sequence of the RKSs with the optimal linear convergence rate.

Then we propose a recursive greedy algorithm for adaptive choice of shifts taking into account non-uniformity of the spectrum. The algorithm is based on an explicit formula for the residual in the frequency domain allowing adaptive shift optimization at negligible cost [5, 6]. The whole procedure only requires to inject some initial rough estimate of the spectral region of the matrix, while further information is automatically generated during the process.

We present numerical experiments, showing that the convergence of a-priori algorithms closely follows the optimal Zolotaryov’s estimate and adaptive approach performs at least as good or better in the case of nonuniform spectral distributions. Also, the adaptive algorithm is less sensitive to quality of the spectral estimate.

Finally, we extend the rational Krylov subspace algorithm from the first order problem to the solution of passive high order dynamical systems

\[ L \left( \frac{d}{dt} \right) u(t) = b(t), \quad u|_{t<0} = 0, \quad L(z) = \sum_{i=0}^{m} A_i z^i, \]
where \( m \in \mathbb{N} \cup \left\{ \infty \right\} \), \( A_i = A_i^* \in \mathbb{R}^{N \times N} \), and \( u(t), b(t) \in \mathbb{R}^N \), \( b|_{t<0} = 0 \).

We show that the reduced equation preserves structure of the original problem and derive a-priori error bound via rational approximation on the nonlinear numerical range of \( L \), defined as \( \left\{ -z \in \mathbb{C} : v^* L(z) v = 0 \text{ for some nontrivial } v \in \mathbb{C}^N \right\} \). This bound shows, that a-priori shift selection approach via the Zolotaryov problem can be extended from the first order problem to the high order dynamical systems, just by using the nonlinear numerical range instead of the linear one. The obtained bounds and shift selection approach can be applied to other known structure preserving algorithms [1, 2, 7].

We compare the obtained bound and shift selection approach with numerical results for the fractional order diffusion problem arising in the solution of the dispersive Maxwell’s system [8].

References


Preconditioners Based on Strong Components

Iain Duff and Kamer Kaya

Abstract

We propose a new method for constructing a preconditioning matrix $M$ to accelerate the solution of the system

$$Ax = b,$$

when using Krylov-based iterative methods. The coefficient matrix $A$ is large and sparse and we assume it is irreducible, that is it cannot be permuted to block triangular form. If it is reducible, then we will apply our algorithms to the irreducible blocks on the diagonal.

The proposed method is based on a hierarchical decomposition of the associated digraph into strongly connected subgraphs. It permutes the rows and columns of the original matrix $A$ and obtains a block triangular preconditioning matrix containing a subset of the nonzeros of $A$ where the maximum size of a diagonal block is smaller than a desired value.

Let $G = (V, E)$ be a strongly connected digraph with $n$ vertices and $m$ weighted edges. A hierarchical decomposition of $G$ into its strong components can be defined in the following way. Let $\sigma_0$ be a permutation of the edges. For $1 \leq i \leq m$, let $\sigma_0(i)$ be the $i$th edge in $\sigma_0$. Let $G_0 = (V, \emptyset)$ be the graph obtained by removing all the edges from $G$. Consider that edges are added one by one to $G_0$ in the order determined by $\sigma_0$. Let $G_i = (V, \{\sigma(j) : 1 \leq j \leq i\})$ be the digraph obtained after the addition of the first $i$ edges. Initially in $G_0$, there are $n$ strong components, one for each vertex, and, during the edge addition process, the strong components gradually coalesce until there is only one. The hierarchical decomposition of $G$ into its strong components with respect to the edge permutation $\sigma_0$ shows which strong components are formed in this process hierarchically. Note that a strong component in a hierarchical decomposition is indeed a strong component of some digraph $G_i$ although it is only a strong subgraph of $G$.

A hierarchical decomposition can be represented with a hierarchical decomposition tree $T$, whose leaf nodes correspond to the vertices in $V$, non-leaf nodes correspond to edges in $E$ that create strong components during the process, and subtrees correspond to the decomposition trees of the strong components that form as the process proceeds. This is better described by diagrams which can be seen in our eponymous report on http://www.cerfacs.fr/6-26857-Technical-Reports-2010.php.

Given a digraph $G = (V, E)$ and a permutation $\sigma_0$, the hierarchical decomposition tree $T$ can be obtained by first constructing $G_0$ and executing Tarjan’s strong component algorithm (SCC) for each internal digraph $G_i$ obtained during the edge addition process. This would be an $O(mn + m^2)$ algorithm since $1 \leq i \leq m$ and the cost of SCC is $O(n + m)$. It would thus be prohibitive for large graphs. To obtain $T$ in a more efficient way, Tarjan first proposed a recursive algorithm of complexity $O(m \log^2 n)$ that he later improved to have complexity $O(m \log n)$ [4].

The first step in our proposed algorithm for creating $M$ is a modified version of Tarjan’s algorithm, HD. Our first modification allows us to have non-distinct edge weights, that is matrices with the same value in different positions. Another modification to the HD algorithm is that we do not want to accept any blocks larger than a predefined value, $m_{bs}$, and this enables us to stop the recursion earlier than in the original algorithm. We use this modified algorithm to obtain a block triangular matrix $M$ where the strong components correspond to the blocks on the diagonal of $M$. To the best of our knowledge, this is the first work that uses Tarjan’s algorithm for preconditioning purposes. After we obtain the block triangular form from this modified HD we then see whether any blocks can be merged and finally use a greedy algorithm to order the blocks on the diagonal so that most of the entries are in the upper triangular part.
The main parameters of our algorithm are the ordering $\sigma_0$ and the largest block size $mbs$. Before we use algorithm HD we first scale and permute the matrix using MC64 [1], which we do also for our experiments on other preconditioners.

We compare our algorithm SCPRE with another block preconditioner XPABLO [3] and a MATLAB version of the industry-standard ILUT on sets of matrices from circuit (left half) and device (right half) simulations from the University of Florida sparse matrix collection. We show below a very abbreviated table from the results in [2]. In the table, we give the number of iterations (with the least in bold font) and the relative memory requirement (in the second line for each matrix). Although this is a much reduced set from our technical report, the results are representative in the sense that the conclusions from this set match those from the larger set.

<table>
<thead>
<tr>
<th>Matrix</th>
<th>XPABLO</th>
<th>ILUT</th>
<th>SCPRE</th>
<th>Matrix</th>
<th>XPABLO</th>
<th>ILUT</th>
<th>SCPRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G_2_{circuit}$</td>
<td>727</td>
<td>89</td>
<td>642</td>
<td>$2D_{27628}$</td>
<td>1.85</td>
<td>2.58</td>
<td>2.22</td>
</tr>
<tr>
<td>circuit_3</td>
<td>572</td>
<td>3</td>
<td>14</td>
<td>$3D_{28984}$</td>
<td>232</td>
<td>-</td>
<td>98</td>
</tr>
<tr>
<td>bircuit</td>
<td>1.25</td>
<td>2.16</td>
<td>1.40</td>
<td>ibm_matrix_2</td>
<td>2.82</td>
<td>1.98</td>
<td>2.65</td>
</tr>
<tr>
<td>ckt11752_dc_1</td>
<td>1.32</td>
<td>1.10</td>
<td>1.38</td>
<td>matrix_9</td>
<td>5.39</td>
<td>23.24</td>
<td>5.12</td>
</tr>
<tr>
<td>mult_dcop_01</td>
<td>1.02</td>
<td>2.55</td>
<td>1.32</td>
<td>wang3</td>
<td>4.96</td>
<td>37.32</td>
<td>2.27</td>
</tr>
<tr>
<td></td>
<td>1.04</td>
<td>22.48</td>
<td>0.86</td>
<td></td>
<td>2.42</td>
<td>8.02</td>
<td>3.85</td>
</tr>
</tbody>
</table>

For circuit simulation problems, ILUT and SCPRE converge for all matrices in this set. XPABLO fails to converge for bcirc and ckt11752_dc_1 and so SCPRE is clearly the best block based preconditioner on this set of matrices. Although ILUT requires significantly fewer iterations on $G_2_{circuit}$ and ckt11752_dc_1, in both cases it requires more memory. However, for $G_2_{circuit}$, if we increase $mbs$ from 1000 to 5000, the number of iterations drops to 95 and our relative memory requirement increases to 6.10 and, for ckt11752_dc_1, by increasing $mbs$ to only 3000 we require only 11 iterations with a relative memory cost of only 1.45. Thus we feel we can recommend using SCPRE for circuit simulation matrices especially when the amount of memory to store the preconditioner is the main concern.

For the device simulation matrices in the right-hand side of the table, the block based preconditioners are far more robust on this set with convergence for all the test matrices. We therefore feel that we can recommend SCPRE as the preconditioner for the device simulation matrices.

To balance these excellent results, we show in our paper [2] that ILUT outperforms both block approaches on matrices from CFD applications. Further research is needed to understand the effect of structure in determining the best approach.

References


On Arbitrary Convergence Behavior of the Arnoldi Method

Jurjen Duintjer Tebbens and Gérard Meurant

Abstract

In the early days of the computer it was believed that the Lanczos method is of no practical use for solving algebraic Hermitian eigenvalue problems. The observed rapid loss of orthogonality between basis vectors in finite precision prevents the method, in general, from finding an acceptable approximation of the spectrum. The opinion on the value of the Lanczos method for practice changed with the appearance of the PhD thesis of Paige [8], who recommended its use for efficient approximation of only a few eigenvalues of a large sparse matrix. Paige showed not only that loss of orthogonality poses no problem for evaluating the quality of computed eigenvalue approximations (Ritz values), but that, on the contrary, convergence of Ritz values goes hand in hand with loss of orthogonality (see [8], or also [7]). An important result needed to prove this is the persistence theorem, which states that once a Ritz value $\theta$ is at a certain distance $\delta$ from an eigenvalue, all Lanczos iterations to come will produce a Ritz value at a distance of at most $\delta$ from $\theta$. In the terminology of [8], the Ritz value $\theta$ is stabilized to within $\delta$. The persistence theorem is a consequence, among others, of the fact that Ritz values produced at subsequent Lanczos iterations interlace, i.e. between every pair of Ritz values there lies a Ritz value of the previous iteration.

Generalizations of the Lanczos method for non-Hermitian matrices basically consist of two classes of methods. Methods of the Bi-Lanczos type exploit short recurrences to build a pair of bi-orthogonal bases for the involved Krylov subspaces and Arnoldi type methods compute a single, in exact arithmetics orthogonal, basis with long recurrences. Methods of both classes are nowadays widely used for the solution of large sparse eigenvalue problems, especially block or band variants in combination with a robust deflation strategy. Because of the long recurrences, Arnoldi methods are in practice restarted; a popular restart strategy is implicit restarting [11], which is the main feature of one of the most efficient software packages for large non-Hermitian eigenvalue problems, ARPACK [5].

Recent convergence results of the Arnoldi method focus on properties of restarted versions, in particular versions using implicit restarts [2]. In this talk we are interested in more basic, theoretical properties of the unrestarted Arnoldi method. We are mainly motivated by the questions: Seen the ability of the Arnoldi method to find accurate approximations of eigenvalues for a large variety of problems, can a result on stabilization of Ritz values as in the Lanczos method be proven? If not so, what is the worst behavior one can expect from the Arnoldi method?

For the Lanczos method, the worst behavior was described by Scott in [9]. Given a Hermitian matrix of order $n$ with the eigenvalues $\lambda_1 < \lambda_2 < \cdots < \lambda_n$, he constructed a perverse starting vector $v$ such that the Ritz values at the last Lanczos iteration are

$$\lambda_1 + \lambda_2 \frac{1}{2}, \lambda_2 + \lambda_3 \frac{1}{2}, \cdots, \lambda_{n-1} + \lambda_n \frac{1}{2}.$$ 

That is, convergence is postponed until the very last iteration and the final Ritz values are as far from the eigenvalues as can be allowed by the interlacing property. For the formulation of an analogue of this result for the Arnoldi method one first has to find a generalized interlacing property for the non-Hermitian case. Some work on this topic has been done by both numerical analysts and scientists from other area’s like Lie Algebra, see e.g. [10, 6]; in [6] one finds a geometrical interpretation of an interlacing property for the Ritz values generated with normal matrices. This interlacing property can be generalized to diagonalizable matrices with considerable effort but the
general case poses serious difficulties. We will explain this briefly in our talk and then we present our main result: Given any set of complex numbers
\[ \theta_1^{(1)}, \theta_1^{(2)}, \theta_2^{(2)}, \ldots, \theta_1^{(n-1)}, \ldots, \theta_{n-1}^{(n-1)}, \theta_1^{(n)}, \ldots, \theta_n^{(n)}, \]
there exists a whole class of matrices with starting vectors such that when the Arnoldi method is applied to members of the class, it generates at the \( k \)th iteration the Ritz values \( \theta_1^{(k)}, \ldots, \theta_k^{(k)} \), \( k = 1, \ldots, n \). In addition, we show how to construct such matrices and starting vectors by giving a full parametrization of the class that generates \( n(n+1)/2 \) prescribed Ritz values.

Thus we have shown that the Arnoldi method can exhibit arbitrary convergence behavior. As a consequence, a persistence theorem is not possible for the Arnoldi method and no stabilization of Ritz values can be guaranteed. It also follows that there is no interlacing property for the Ritz values generated by general non-Hermitian matrices. Our result is similar in spirit to the results of Arioli, Greenbaum, Pták and Strakoš showing that the GMRES method can generate any convergence curve with any spectrum [4, 3, 1]. While the GMRES method is closely related to the Arnoldi method, one may suspect our result to be a straightforward consequence of [4, 3, 1]. This is not the case, although we do exploit the theory developed in [4, 3, 1]. In fact, our result implies that GMRES can generate a prescribed convergence curve not only with any spectrum but even with any distribution of the \( n(n+1)/2 \) Ritz values produced during the iterative process, except for one Ritz value at each iteration.

References

Randomized Algorithms for Communication-Optimal SVD and EIG

Grey Ballard, James Demmel, Ioana Dumitriu

Abstract

Randomization has long been a useful technique in the study of algorithms, whether for “average” or “typical” performance studies, or for understanding the good behavior, in practice, of algorithms known to have terrible worst-case bounds in theory. More recently, however, randomization has been steadily growing in importance as a very powerful tool in low-rank matrix approximation. For a good survey paper, see [5]. While the algorithms catalogued in [5] are very effective at dealing with low-rank matrices, and can be used as part of actual SVD/EIG calculations for low-rank matrices, the techniques cannot be applied directly to the calculation of the spectrum or singular values of a generic, large matrix. This is the problem that our algorithms in [3] address for the symmetric, non-symmetric, and generalized eigenproblems, as well as for the singular value problem.

The running time for an EIG/SVD algorithm for a very large matrix, which does not fit into the fast memory of a processor, depends on two costs: arithmetic and communication. The latter represents the data-passing between different layers of the memory hierarchy (in the sequential case) or between the various processors in a network (in the parallel one). In particular, given the “costs” \( \gamma \) associated to each arithmetic operation executed, \( \beta \) associated to each “word” transmitted (\( \beta \) can be expressed as the inverse of the bandwidth), and \( \alpha \) (the latency) associated to each message transmitted, the “total cost” is then expressible as

\[
\text{Cost (or running time)} = \gamma \cdot (\text{# arithmetic operations}) + \\
\beta \cdot (\text{# words transmitted}) + \\
\alpha \cdot (\text{# messages transmitted})
\]

We have developed in [4] and further improved in [3] a set of randomized algorithms for computing the spectrum and singular values of a very large generic matrix (by “generic”, we mean that we do not expect the spectrum or singular values to have accumulation points) or pencil. These algorithms are based on the algorithm first presented in [1]. We showed in [3] that these algorithms minimize, in a big-Oh sense, all arithmetic and communication costs (both bandwidth and latency), in a cache-oblivious way, achieving the lower bounds on communication proved (for the case of \( O(n^3) \) algorithms) in [2]. In addition, randomization brings a better stability guarantee than in the case of classical algorithms.

To minimize communication, the algorithms do more arithmetic (by a multiplicative factor depending on the problem and error tolerance) than classical algorithms. Nevertheless, given that on current hardware, \( \beta \) and \( \gamma \) greatly exceed \( \alpha \), and technology trends indicate that this processor-memory gap is growing exponentially over time, it is becoming imperative to devise algorithms that minimize communication—sometimes by trading off arithmetic for communication, as we do here.

As an aside, we showed in [4] that our randomized algorithms are as fast and stable as the black-box fast matrix multiplication algorithm (faster than \( O(n^3) \); e.g., Strassen) they use. New developments seem to indicate that Strassen-like algorithms for fast matrix multiplication are also communication-avoiding, and reach the same lower bounds proved in [2].

The key component in the algorithm (discussed throughout the rest of this abstract in its non-symmetric form) is a single step of Divide-and-Conquer. By means of a generalized randomized
rank-revealing factorization (introduced and analyzed in [4], and generalized in [3] to work in the case of any product of matrices and inverses, without explicitly constructing the product), the spectrum is divided into two parts. These parts are defined as the “outside” and the “inside” of a region in the plane, which can be constructed from a Gershgorin circle by repeated intersections with lines and circles. In order to drive the eigenvalues apart, the algorithm performs “disguised” repeated squaring (which is achieved without calculating inverses or any other explicit matrix functions).

At the end of this stage, the algorithm splits the problem into two smaller ones, by producing a block-diagonal decomposition of the matrix $A$ and outputting

$$
\hat{A} = Q^H AQ = \begin{pmatrix} A_{11} & A_{12} \\ E_{21} & A_{22} \end{pmatrix};
$$

the matrices $A_{11}$, respectively, $A_{22}$ have all their eigenvalues outside, respectively, inside the region. The dimensions of the subblocks are chosen so as to minimize $||E_{21}||_1/||A||_1$, and the orthogonal matrix $Q$ represents the deflating subspace corresponding to the “inside” eigenvalues.

The algorithm is then repeated on $A_{11}$ and $A_{22}$, with new choices for splits.

In addition to the basic, “low-level” step of Divide-and-Conquer, we have also given a “high-level” strategy for how it is best to choose the lines and circles for the splits, and we have shown that progress is made with high probability. We describe so called “pessimal” cases, and explain why some notoriously ill-conditions problems are handled with ease by randomizing the choice of splits. The only truly pessimal case is that of a tightly packed pseudospectrum, in which case we output a convex hull for it, in the nonsymmetric case (conventional algorithms output a uniform sample of the pseudospectrum), or a small interval containing all the eigenvalues, in the symmetric one.

To summarize, with high probability, each step of Divide-and-Conquer produces a good split, with only a few iterations of disguised repeated squaring. The good split is guaranteed; “with high probability” here refers to the number of iterations being small.

If given more time, I will present this algorithm in the context of the larger class of randomized algorithms mentioned in the beginning of the abstract; otherwise, I will focus on the double use of randomization in the algorithm (both in the low-level and high-level steps), and detail its benefits.

References


Needle-like Triangles, Matrices and Lewis Carroll

Alan Edelman, and Gilbert Strang

Abstract

Triangles

Kahan, in a wonderful and widely known investigation, studied the numerical computation of the area of a triangle given the edge lengths $a, b,$ and $c$. The main message of Kahan’s work is well known, that the side lengths should be sorted as $a \geq b \geq c$ and then area should be computed exactly without removing parentheses as:

$$K = \frac{1}{4} \sqrt{((a + (b + c))(c - (a - b))(c + (a - b))(a + (b - c)))}.$$

The formula is also accurate when the sides are presumed to have rounding errors and the area is not an ill-conditioned function of the side lengths. Well-conditioning occurs when the triangle is needle-like. No formula can be accurate if the data is presumed to have rounding errors and the area is an ill-conditioned function of the side lengths.

Matrices

The numerical computation of the smallest singular value of a $2 \times 2$ matrix can be fairly accurate as long as the matrix is not too badly conditioned. High relative accuracy is also available so long as the smaller singular value is not an ill-conditioned function of the matrix entries. Well conditioning occurs when the matrix is diagonal or very nearly so.

Shape Theory

“Shape Theory” brings together theoretically the two problems above so that they are no longer simply “analogies.” We discuss how we rediscovered and then stumbled upon the theory, the last interest of the first professor of mathematical statistics at Cambridge University, David Kendall. Along the way we will revisit Lewis Carrol’s question: Are most random triangles acute or obtuse?
Computing Low-Rank Approximations of Sparse Tensors using Krylov Methods, and Applications in Information Sciences

_Lars Eldén, and Berkant Savas_

Abstract

Analyses of data organized as matrices are ubiquitous and well supported by theory and algorithms. For large and sparse matrices the main class of algorithms is Krylov methods. In many applications data are organized in more than two categories, and it is often unnatural to reorganize the data as a matrix. Thus there is a need for theory and algorithms for tensor computations. In many applications, especially in information sciences, the tensors are large and sparse.

Recently we have discovered how to generalize matrix Krylov methods for the computation of low-rank approximations of tensors. We present different alternative ways of applying the Krylov idea. A direct generalization of the Golub-Kahan and Arnoldi procedures leads to a method that can be called a minimal Krylov recursion. As the combinatorial structure is much richer than in the matrix case, one may generate many more vectors than for matrices; this leads to a maximal recursion. A maximal recursion admits a Krylov factorization of the tensor. We also describe an optimized recursion that generates as few vectors as the minimal one, but has better approximation properties.

A low-rank approximation of a matrix can be refined, e.g. for the computation of eigenvalues, using a restarted Arnoldi method or a Krylov-Schur algorithm. We present a generalization for tensors.

A number of examples are given, where the procedures are applied to large and sparse tensors. The different methods are compared with regard to computational efficiency and accuracy. The use of tensor methods in a couple of areas of information science is illustrated.
Linear Algebra Problems Arising from Parameter-Dependent Partial Differential Equations

Howard C. Elman

Abstract

Mathematical models based on partial differential equations often have parameters associated with them. Such parameters may correspond to material properties, geometry or boundary conditions associated with the model, or when some of these quantities (for example, diffusion coefficients) are uncertain and treated as random variables. In these situations, there are many places where linear algebra algorithms play a role in solution of the underlying model, and it is important that these be performed efficiently in order to obtain accurate solutions in an efficient way. These include:

1. Use of multigrid methods to solve the large-scale algebraic systems that arise from discretization of models. In particular, for problems in which the parameters are random quantities, there is a variety of ways in which multigrid can be used to obtain solutions and statistical quantities of comparable accuracy, and the costs of the different strategies may vary significantly.

2. Reduction of the dimensionality of large-scale discrete models using sampling techniques. This is done by solving the parameter-dependent system for a small number of sample values of the parameter set, and then using a projection of the model into the space of sampled solutions to estimate solutions for other parameters. Strategies for choosing the samples effectively entail optimization of an estimate for error over a wide choice of samples, and simulation of many parameters entails solution of many small or moderate-sized systems of equations.

3. Representation of multiple-parameter models using a smaller number of parameters. This can be done by taking a collection of sample values of the parameters, constraining the parameter-dependent terms at these values, and then using so-called principle components analysis or proper orthogonal decomposition to reduce the dimension of the parameter space. This essentially entails construction of singular value decompositions and using a small number of the singular vectors (components) corresponding to the largest singular values to reduce the number of parameters.

We outline the algebraic computations associated with each of these strategies and assess the comparative costs of using them for numerical models of diffusion.

Representative examples of studies of the first two approaches are


The third idea is widely used for parameter-dependent models. Its use in the setting of PDEs is described in http://icms.org.uk/downloads/uq/Matthies.pdf
The Stability of GMRES Convergence with Applications to Inexact Preconditioning

Mark Embree, Ronald Morgan, Josef Sifuentes, Gilbert Ymbert

Abstract

How does the convergence of the GMRES algorithm for linear systems change when the coefficient matrix is perturbed? Suppose the \( k \)th iteration of GMRES applied to the matrix \( A \) and starting vector \( b \) yields a residual \( r_k = p(A)b \), from which we wish to draw conclusions about the GMRES residual for the matrix \( A + E \) with the same starting vector. The natural approach of explicitly comparing \( p(A) \) to \( p(A + E) \) would require explicit knowledge of harmonic Ritz values (or, equivalently, the coefficients of \( p \)), a degree of insight well beyond our present understanding; the potential instability of the spectrum of \( A \) to small perturbations adds additional complication.

In this talk, we shall describe a different approach: write \( p(A) \) and \( p(A + E) \) in their Cauchy integral forms and apply standard perturbation theory for the resolvent (a technique perhaps first used by Rinehart in 1956 [9, 4]). Such analysis leads to the following theorem [6].

**Theorem.** Let \( r_k = p(A)b \) and \( \rho_k \) denote the residuals produced by \( k \) iterations of GMRES applied to \( Ax = b \) and \( (A + E)x = b \), with \( p(0) = 1 \). Then for all \( \delta > \|E\| =: \varepsilon \),

\[
\frac{\|\rho_k\|}{\|b\|} \leq \frac{\|r_k\|}{\|b\|} + \left( \frac{\varepsilon}{\delta - \varepsilon} \right) \left( \frac{L_\delta}{2\pi\delta} \right) \sup_{z \in \sigma_\delta(A)} |p(z)|
\]

and

\[
\frac{\|\rho_k\|}{\|b\|} \leq \left( \frac{\delta}{\delta - \varepsilon} \right) \left( \frac{L_\delta}{2\pi\delta} \right) \min_{\deg(q) \leq k} \sup_{q(0) = 1} \frac{1}{z \in \sigma_\delta(A)} |q(z)|,
\]

where \( L_\delta \) denotes the length of boundary of \( \sigma_\delta(A) \), the \( \delta \)-pseudospectrum of \( A \).

The pseudospectrum arises naturally in these bounds due to the resolvent term in the Cauchy integrals, giving a ready way to handle the potential instability of the spectrum of \( A \) to perturbations; see also [1]. (For examples of pseudospectra-based bounds for GMRES convergence, see [8, §26].)

While the proposed analysis may have some utility for standard GMRES (e.g., relating to inexact matrix-vector products [7] or low-rank updates to \( A \) in optimization), we believe the most compelling applications arise in the context of preconditioning. Often an exactly-applied preconditioner leads to a coefficient matrix \( M^{-1}AP^{-1} \) with special spectral structure that ensures GMRES exhibits some appealing convergence. Practical considerations prevent the exact application of the preconditioner, thus altering the promised GMRES behavior. To what extent is convergence compromised?

We shall discuss several specific examples. Sifuentes has conducted a careful analysis of inexact deflation preconditioning [5]. Another interesting practical example arises from the Murphy–Golub–Wathen preconditioner for saddle point problems [3]. Exact preconditioning delivers a coefficient matrix with three eigenvalues, giving exact convergence in three iterations at the expense of exactly inverting several matrices (independent of the sensitivity of the three eigenvalues to perturbations). We will show how our analysis provides insight into the convergence of GMRES when the constituent matrices are inexactly inverted, now dependent on the spectral instability of the exactly preconditioned system.
References


Multilevel Krylov Method for the Biharmonic Equation

Yogi Ahmad Erlangga

Abstract

In this talk, we discuss the application of the multilevel Krylov method [1], MK in short, to the linear system of equations

\[ Ax = g, \quad A \in \mathbb{R}^{n \times n}, \tag{1} \]

associated with the biharmonic equation

\[ \Delta \psi = f, \tag{2} \]
\[ \psi - \Delta \phi = 0, \tag{3} \]

with two sets of boundary conditions (bc): (i) \( \phi = \Delta \phi = 0 \) and (ii) \( \phi = \partial \phi / \partial \eta = 0 \), where \( \eta \) is the directional vector normal to the boundaries. With the first set of boundary conditions, \( A \) is a \( 2 \times 2 \) block triangular matrix, which allows a sequence of block solves, starting from (2). The second set of boundary conditions forms a coupled block matrix that requires a simultaneous solve of the linear system. In either case, we shall solve (1), simultaneously.

Let

\[ P_A = I - AZ\mathcal{E}_A^{-1}Z^T + \lambda_n Z\mathcal{E}_A^{-1}Z^T, \tag{4} \]

called the shift operator, with \( \mathcal{E}_A = Z^TAZ \in \mathbb{R}^{r \times r} \) the Galerkin matrix, \( Z \in \mathbb{R}^{n \times r}, r < n \), the prolongation matrix of rank \( r \), and \( \lambda_n = \max\{|\lambda| : \lambda \in \sigma(A)\} \), where \( \lambda \) is the eigenvalue of \( A \). A multilevel Krylov method is obtained if the preconditioned linear system

\[ P_A Ax = P_A g, \tag{5} \]

is solved by a Krylov method, with the Galerkin system associated with \( \mathcal{E}_A^{-1} \) solved via a similar preconditioned system as above with the appropriate shift operator. For (5), if \( \sigma(A) = \{\lambda_1, \ldots, \lambda_n\} \), with \( 0 < \lambda_i \leq \lambda_j, i < j \), and if \( \text{col} Z = \{v_1\}_{i=1}^r \) are the eigenvectors associated with the \( r \) smallest eigenvalues of \( A \), then \( \sigma(P_A A) = \{\lambda_{r+1}, \ldots, \lambda_n\} \). As the spectrum of \( P_A A \) is better clustered than \( A \), a faster convergence can be expected. For practicality, it suffices to use approximate vectors to \( v_i \).

We set \( Z = \begin{bmatrix} Z & 0 \\ 0 & Z \end{bmatrix} \), where \( Z \in \mathbb{R}^{m \times s} \) of rank \( s < m \). With bc (i), if \( \sigma(L) = \{\mu_1, \ldots, \mu_m\}, m = \frac{1}{2}n \) is the spectrum of the discrete Laplacian \( L \in \mathbb{R}^{m \times m} \), then \( \sigma(A) = \sigma(L) \). Moreover, let

\[ P_L = I - LZ\mathcal{E}_L^{-1}Z^T + \mu_m Z\mathcal{E}_L^{-1}Z^T, \quad \mu_m = \max\{|\mu| : \mu \in \sigma(L)\} \tag{6} \]

with \( E_L = Z^T L Z \), and \( Z \in \mathbb{R}^{m \times s}, s < m, s = \frac{1}{2} r \). We have that \( \sigma(P_A A) = \sigma(P_L L) \), with \( P_L A \) representing an MK method applied to the Laplacian \( L \). With the same spectrum, MK applied to the biharmonic equation will converge as fast as MK for the Poisson equation. Furthermore, if the Galerkin matrix is perturbed by dropping its off-diagonal blocks, we have that \( \sigma(\bar{P}_A A) = \sigma(P_L L) = \sigma(\bar{P}_A A) \), where \( \bar{P}_A \) is (4) with \( \mathcal{E}_A \) replaced by the perturbed one. Hence, we can also in

\[ \text{Y.A. Erlangga, On the multilevel Krylov method for the biharmonic equation, 2010.} \]
this case expect the same fast convergence as MK for the Poisson equation. With the coarse-grid block matrix now completely decoupled, we can in principle use any efficient local/block method to solve the coarse-grid system.

The case with bc (ii) requires a more involved analysis. One point is that as the associated block matrix $A$ is strongly coupled, a similar perturbation to the Galerkin matrix leads to a spectrum $\sigma(\tilde{P}_A)$ that differs significantly from $\sigma(P_L A)$, and hence can not be employed. The general spectral properties of MK as presented in [1], however, hold.

We shall present in the talk, convergence results with MK, some of which are shown in Tables 1 and 2 for the case with bc (i) and bc (ii), respectively. The two-level Krylov method is the ideal scenario, in which $E_A$ is inverted exactly. This in principle leads to the “best” convergence one can attain for a given MK setup. With MK(4,2), the system associated with $E_A$ is solved approximately and iteratively by applying a similar operator to (4) on the coarse level recursively, with 4 Krylov (in this case GMRES) iterations at the second level, and 2 at the third or higher level. For the case with bc(i) and with the full Galerkin $E_A$, the convergence of MK(4,2) is grid independent, and is similar to MK for the Poisson equation (Table 1). The same results are also obtained with $\tilde{E}_A$. For the case with bc(ii), the MK(4,2) convergence deteriorates as the mesh size decreases. A grid-independent convergence can however be recovered by adding extra work on the coarse levels, as in MK(4,4).

Table 1: Number of MK iterations for the biharmonic equation with bc (i) and full Galerkin. The same results are obtained with the perturbed Galerkin. $N$ is the # of grid points to discretize $\Delta$.

<table>
<thead>
<tr>
<th>$N$</th>
<th>$16^2$</th>
<th>$32^2$</th>
<th>$64^2$</th>
<th>$128^2$</th>
<th>$256^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-level</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>MK(4,2)</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 2: Number of multilevel Krylov iterations for the biharmonic equation with bc (ii).

<table>
<thead>
<tr>
<th>$N$</th>
<th>$16^2$</th>
<th>$32^2$</th>
<th>$64^2$</th>
<th>$128^2$</th>
<th>$256^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-level</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>MK(4,2)</td>
<td>13</td>
<td>12</td>
<td>14</td>
<td>25</td>
<td>35</td>
</tr>
<tr>
<td>MK(4,4)</td>
<td>13</td>
<td>12</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
</tbody>
</table>

We note that such a fast, grid-independent convergence is obtained with $Z$ based on an aggregation technique adopted from algebraic multigrid methods, and with $\lambda_n$ estimated via Gershgorin’s theorem.

References

Krylov Subspace Methods for Geoelectrical Exploration Problems

Michael Eiermann, Oliver G. Ernst, and Stefan Güttel

Abstract

Geoelectrical exploration refers to the measurement of electromagnetic fields due to natural or synthetic sources to infer the distribution of electromagnetic conductivity in the subsurface. Applications include the detection of hydrocarbons, mineral deposits and unexploded ordnance as well as the monitoring of CO2 sequestration. The numerical simulation of a widely used exploration method leads to parameter estimation problems for determining the conductivity function $\sigma$ in time-dependent quasi-static Maxwell’s equation

$$\mu_0 \sigma \dot{E} = -\nabla \times (\nabla \times E) - \mu_0 J_i$$

with electric field vector $E$, constant magnetic permeability $\mu_0$ and impressed source current $J_i$ on a bounded computational domain $\Omega \subset \mathbb{R}^3$. With suitable initial conditions at time $t = t_0$ and after discretization in space using Nédélec finite elements, one obtains the linear ordinary initial value problem

$$M \dot{u} = Ku, \quad u(t_0) = u_0,$$

for the vector $u$ of coefficients of the finite element representation of the electric field, where $M$ and $K$ are the large and sparse finite element mass and stiffness matrices, respectively. Its solution can be represented in terms of the action of the matrix exponential $v(t) = e^{(t-t_0)A}v_0$ on an initial vector $v_0$.

In this talk we present our experiences with solving this initial value problem using Krylov subspace approximations for the matrix exponential operator. We compare traditional time-stepping techniques with polynomial and rational Krylov subspace iteration. For the latter, the dimension of the Krylov space required for a given level of accuracy remains bounded independently of the spatial mesh size. Each rational Krylov step, however, requires solving a linear system with $K$. We demonstrate how this can be achieved efficiently using specially designed multigrid methods.

We also discuss the stationary problem arising from the frequency-domain analogue of equation (1) (cf. [1]). In this case, the task is the evaluation of the expression

$$f(\omega) = R^\top (K + i\omega M)^{-1} q,$$

where the restriction operator $R$ maps the full field to the values of a small number of measurement functionals, for a large number of discrete frequencies $\omega$. We achieve this using a Krylov-subspace model reduction method in which the Krylov subspace approximation of the right hand side of (2) is used as an inexpensive surrogate. Drawing on results of the recent Ph.D. thesis of Stefan Güttel [2], we discuss the tradeoffs between polynomial and different variants of rational Krylov subspace methods as well as optimal parameter choices for these algorithms.

References


Numerical Linear Algebra in POD for Model Order Reduction of Steady Aerodynamic Applications

Heike Fassbender, and Alexander Vendl

Abstract

A major part of the development costs of an aircraft is incurred by wind-tunnel testing and actual flight tests which are needed in order to obtain aircraft certification. In case a problem is detected at this stage of the development it is extremely expensive to modify the design of the aircraft in order to deal with the problem. Hence it is of interest to complement the wind tunnel and flight tests by increased use of numerical simulations.

The numerical simulation of aircraft aerodynamics requires the solution of a suitable mathematical model taking all relevant physical effects into account. For an industrial aircraft configuration, each simulation may take several hours on a parallel computer using hundreds of cores. In order to fully simulate the behavior of an aircraft it is estimated that up to 20,000,000 such simulations are required (using engineering experience for current configurations and technologies, still about 100,000 simulations are necessary). In order to be able to achieve this in feasible time one can make use of model reduction techniques. Model reduction seeks to replace the large-scale mathematical model with a small-scale one which is much faster to simulate (in a few seconds or minutes), but still captures the relevant characteristic behavior of the aircraft.

Assuming that a suitable full order model describing the proper flow physics is available, using a computational fluid dynamics (CFD) solver, flow solutions can be computed for certain parameter settings (describing certain flight conditions). Using these so-called snapshots, low dimensional models can be generated using the proper orthogonal decomposition (POD) ansatz. These reduced-order models suffer from a number of problems; in particular, they are only valid for parameter settings very close to those used to compute the snapshots.

Several modifications of the basic POD method have been proposed in the literature. We have implemented and tested the gappy POD, missing point estimation and discrete empirical interpolation for various questions arising in the context of aircraft simulation.

E.g., the so-called Gappy POD was implemented to be used for the purpose of data fusion. This approach allows combining data obtained from numerical simulations with partial experimental data to reconstruct an entire flow field. The approach has been demonstrated for fusing gappy wind tunnel data of an industrial aircraft configuration with CFD data. It can also be used for design purposes such as inverse airfoil design.

In this talk we will highlight some of the numerical linear algebra aspects related to our POD implementations. We will illustrate the benefits of using POD models for aerodynamics by presenting some data fusion examples.
LSMR: An Iterative Algorithm for Least-squares Problems

David Chin-lung Fong, Michael Saunders

Abstract

For nearly 30 years, LSQR has been the standard iterative solver for large rectangular systems $Ax \approx b$. It is analytically equivalent to symmetric CG on the normal equations, and it reduces $\|r_k\|$ monotonically, where $r_k = b - Ax_k$ is the $k$-th residual vector.

LSMR is equivalent to applying MINRES to the normal equations, so that $\|A^T r_k\|$ decreases monotonically. In practice we observe that $\|r_k\|$ and the Stewart backward error $\|A^T r_k\|/\|r_k\|$ are also monotonic, and the backward error is usually very close to optimal. Thus if iterations need to be terminated early, it is safer to use LSMR.

Both methods are based on the Golub-Kahan bidiagonalization process (a short-term recurrence for generating vectors $u_k$ and $v_k$). Experiments show that if the vectors $v_k$ are reorthogonalized, then the vectors $u_k$ remain orthogonal, and (almost) vice versa.

Matlab, Fortran 90 and Python implementations of LSMR are available, with local reorthogonalization of $v_k$ as an option (www.stanford.edu/group/SOL/software.html). Plots of various quantities on a range of large test problems illustrate the desirable properties of LSMR.
Reliable Calculation of Numerical Rank, Null Space Bases, Basic Solutions and Pseudoinverse Solutions using SuiteSparseQR

Leslie V. Foster

Abstract

The numerical rank of a matrix can be defined as the number of singular values larger than a specified tolerance tol and calculation of the numerical rank is important in the presence of rounding errors and fuzzy data (Gene Golub and Charles Van Loan, Matrix Computations, third edition, p. 72, 1996). The most accurate algorithms for calculating the numerical rank are based on the singular value decomposition (SVD) but calculation of the SVD can be impractical for large matrices. The routine SPQR from SuiteSparseQR (T.A. Davis, “Algorithm 9xx, SuiteSparseQR: multifrontal multithreaded rank-revealing sparse QR factorization,” submitted to ACM TOMS, 2009, http://www.cise.ufl.edu/~davis/techreports/SPQR/spqr.pdf) is an efficient algorithm for large, sparse matrices and it returns an estimate for the numerical rank that is usually but not always correct. We present a number of routines based on SPQR that improve the accuracy of the numerical rank calculated by SPQR. The new routines calculate estimates of upper and lower bounds of singular values of $A$ and use these estimates to return an error flag that warns the user when the calculated numerical rank may be incorrect. The algorithms reliably determine the numerical rank in the sense that, based on extensive testing with matrices from applications with a common choice for tol, the numerical rank is accurately determined when the warning flag indicates that the numerical rank should be correct. The test set consists of 767 numerically singular matrices contained in the San Jose State University Singular Matrix Database (http://www.math.sjsu.edu/singular/matrices). The matrices in this set come from real world applications or have characteristic features of real world problems and most of the matrices are sparse. In addition to the numerical rank our routines can calculate orthonormal bases for numerical null spaces, approximate pseudoinverse solutions to the least squares problem $\min ||Ax - b||$, where $A$ is an $m \times n$ matrix, as well as basic solutions to these problems. An important feature of our algorithms is that they can be efficient and accurate for large, sparse matrices when the matrix is neither “low rank” nor “high rank”: the numerical rank does not need to be small or close to $n$. For example the routines successfully calculate an orthonormal basis for the numerical null space of a 321671 by 321671 matrix whose numerical null space dimension is 222481.
New Algorithms for Calculating the Distance to Instability and the Distance to a Nearby Defective Matrix

Melina A. Freitag and Alastair Spence

Abstract

We provide new methods for the solution of two matrix nearness problems: First, a new fast algorithm for the computation of the distance of a stable matrix to the unstable matrices is provided. Let $A$ be a stable complex $n \times n$ matrix, that is all its eigenvalues are in the open left half plane. A small perturbation $E$ to the matrix may lead to eigenvalues of $A + E$ crossing the imaginary axis and hence the matrix $A + E$ being unstable. The distance of a matrix $A$ to instability can be described by

$$
\beta(A) = \min_{\|E\|} \{ \eta(A + E) = 0, \ E \in \mathbb{C}^{n \times n} \},
$$

where $\eta(A) = \max\{\Re(\lambda) | \lambda \in \Lambda(A)\}$, or, using the singular value decomposition [4],

$$
\beta(A) = \min_{\omega \in \mathbb{R}} \sigma_{\min}(A - \omega i I).
$$

In [4] a bisection method for computing $\beta(A)$ was introduced. The method provides lower and upper bounds on $\beta(A)$ but requires the solution of a sequence of eigenvalue problems for the $2n \times 2n$ Hamiltonian matrix

$$
H(\alpha) = \begin{bmatrix} A & -\alpha I \\ \alpha I & -A^H \end{bmatrix}
$$

for a positive real $\alpha$. $H(\alpha)$ has a pure imaginary eigenvalue if and only if $\alpha \geq \beta(A)$. Boyd & Balakrishnan [3] proposed a quadratically convergent method for the more general task of finding the $H_\infty$-norm of a transfer function matrix, which reduces to the problem discussed here in the simplest case. These algorithms require the computation of all eigenvalues of $H(\alpha)$ at each step. We provide a new quadratically convergent method, based on finding a two-dimensional Jordan block corresponding to a pure imaginary eigenvalue in (1), that does not require the computation of all eigenvalues of $H(\alpha)$ at each step [5].

Second, a new fast method for the computation of the distance of a matrix to a nearby defective matrix is presented. Let $A$ be a complex $n \times n$ matrix with $n$ distinct eigenvalues. It is a classic problem in numerical linear algebra to find

$$
d(A) = \inf\{\|A - B\|, \ B \text{ is a defective matrix}\}.
$$

We formulate the problem following Alam & Bora [1,2] and reduce it to finding when a parameter-dependent matrix is singular subject to a constraint. The solution is achieved by an extension of the Implicit Determinant Method [6].

Numerical results show the performance of both algorithms for several examples and comparison is made with other methods for the same problem.

References


Theoretical and Numerical Results and Problems in Tensors

Shmuel Friedland

Abstract

In this lecture we survey theoretical and numerical results and problems in tensors that listed below

Generic and typical ranks of 3-tensors

Consider tensors $C^{m \times n \times l}$ where $2 \leq m \leq n \leq l$. Then a generic rank $\text{grank}(m, n, l)$ is the maximal border rank in $C^{m \times n \times l}$. Equivalently, it is the rank of a tensor $T \in C^{m \times n \times l}$, where $mnl$ the entries of $T$ are chosen at random. It is known that $\text{grank}(m, n, l) \leq mnl$ and $\text{grank}(m, n, l) = l$ for $(m - 1)(n - 1) + 1 \leq l \leq mnl$, e.g. [2]. It is conjectured in [2] that $\text{grank}(m, n, l) = \lceil \frac{mnl}{m + n + l - 2} \rceil$ for $2 \leq m \leq n \leq (m - 1)(n - 1)$, and $(m, n, l) \neq (3, 3, 2p + 1)$, for $p = 1, 2, \ldots$. We will discuss briefly this conjecture, some known cases, its numerical proof for $m, n, l \leq 15$ and its relation to fast matrix multiplication.

Next we consider the notion of typical ranks of tensors over $R^{m \times n \times l}$. This is a rank of tensor $T \in R^{m \times n \times l}$ whose entries are chosen at random. In this case one can have more than one typical rank. In practice, so far we know only the cases where there is one or two typical ranks. As I emphasized in [2] the subject of typical ranks belongs to the field of semi-algebraic geometry. In particular, it is shown in [2] that the number of typical ranks in any $R^{m \times n \times l}$ is finite and takes all the integer values in the interval $[\text{grank}(m, n, l), \text{grank}(m, n, l) + k]$. We will discuss some known examples, and the conjecture that $k \leq 1$.

Low rank approximation of tensors

In practical application one needs good low rank approximation of $k$-mode tensors $R^{m_1 \times \cdots \times m_k}$ in general and 3-tensors in particular. Basically, there are two main approaches. Let us explain it for 3-tensors. View $R^{m_1 \times m_2 \times m_3}$ as a vector space $R^{m_1} \otimes R^{m_2} \otimes R^{m_3}$ with the $\ell_2$ norm induced the the standard $\ell_2$ norms in $R^{m_i}, i = 1, 2, 3$. Then one tries to find the best three subspaces $U_i \subset R^{m_i}$ of given small dimension $d_i$ for $i = 1, 2, 3$ such that the projection of $T$ on this subspace is maximal. One of the standard method is a relaxation method, where all but one subspace is kept fixed. A more sophisticated idea is to use singular value decomposition, (SVD), as suggested in [4]. Another natural approach is to use Newton method on the corresponding product of Grassmannian $\times_{i=1}^3 \text{Gr}(d_i, [R^{m_i}])$ [1].

A completely different method is a fast approximation of tensors using CUR decomposition. The origin of this method for matrices is in the paper [6]. It is a very flexible method which can be adopted to tensors [5]. One of the difficulty of this method is a good choice of rows, columns and depth of tensors to be chosen. For matrices, one need to choose a square submatrix of of a given size $k$ with maximal determinant. In the case of symmetric nonnegative definite matrices this problem is a maximization problem of submodular function [3]. We will explain in detail theoretical and numerical aspects of these approaches.
References


Computational Proofs of the Stability of Lyapunov Equations

Andreas Frommer and Behnam Hashemi

Abstract

A prominent task in the analysis of dynamical systems in control theory is to check whether a Lyapunov equation

\[ AX + XA^T = C, \quad \text{where} \ A \in \mathbb{R}^{n \times n}, \ C \in \mathbb{R}^{n \times n} \]  

(1)

has a positive (or negative) definite solution \( X \in \mathbb{R}^{n \times n} \). The purpose of this work is to present a numerical algorithm whose output will be exactly one of the two following statements, where the first one is correct with mathematical certainty:

1. (1) has a symmetric positive definite solution. The algorithm then also provides correct and tight lower and upper bounds for each entry of the solution \( X \).

2. Failure, i.e. we do not obtain any information on whether \( X \) is positive definite or not.

The major ingredient in this algorithm is to compute an interval enclosure \( X \) of \( X \), i.e. a matrix \( X \) whose entries are compact intervals which have been proven to contain the corresponding entries of the solution \( X \). Of course, the aim is to compute as narrow as possible intervals for all of these entries. A classical approach towards this task for a general, unstructured linear system

\[ Px = c, \]

is Krawczyk’s method. Given an approximate solution \( \bar{x} \) of the linear system, computed by some floating point linear system solver, and given an approximate inverse \( R \) of \( P \), again computed by some floating point algorithm, Krawczyk’s method uses machine interval arithmetic (including outward rounding) to check whether

\[ \{ R(c - P\bar{x}) + (I - RP)z : z \in z \} \subseteq \text{int } z \]  

(2)

where \( z \) is an appropriately chosen interval vector. If this is the case, one knows [3] that \( P \) and \( R \) are non-singular, and that the solution of \( Px = c \) is contained in \( \bar{x} + z \).

Denoting \( x = \text{vec}(X) \) and \( c = \text{vec}(C) \), (1) can be written as

\[ Px = c, \quad \text{where} \ P = I \otimes A + A \otimes I \in \mathbb{R}^{n^2 \times n^2}. \]  

(3)

Applying Krawczyk’s method to (3) would result in a computational complexity of \( \mathcal{O}(n^6) \) due to the computation of the approximate inverse \( R \). An important observation is that the complexity can be reduced to \( \mathcal{O}(n^3) \) if \( A \) is diagonalizable, see [2]. Assume that we have computed approximate (left) eigenvectors and corresponding eigenvalues, i.e. we have a matrix \( W \) and a diagonal matrix \( D \) such that

\[ WA \approx DW. \]

Then \( Y = WXW^T \) is the solution of the Lyapunov equation

\[ (WAW^{-1})Y + Y(WAW^{-1})^T = WCW^T. \]  

(4)

Since \( WAW^{-1} \) is (almost) diagonal, the matrix \( I \otimes WAW^{-1} + WAW^{-1} \otimes I \) is almost diagonal, too, and it allows for a diagonal approximate inverse. It is thus possible to formulate a variant of
Krawczyk’s method which computes an enclosure $\mathbf{Y}$ for the solution $\mathbf{Y}$ of (4). Note that $W^{-1}$ is a quantity which is not available exactly in floating point, so this variant of Krawczyk’s method uses a computable interval enclosure of $W^{-1}$.

Since $\mathbf{X} = W^{-1} \mathbf{Y} W^{-T}$, the matrix $\mathbf{X}$ is positive definite if and only if $\mathbf{Y}$ is positive definite. All we know is $\mathbf{Y} \in \mathbf{Y}$, so we have to show that every symmetric matrix contained in $\mathbf{Y}$ is positive definite or, alternatively, that all symmetric matrices from $\mathbf{X} := W^{-1} \mathbf{Y} W^{-T}$ are positive definite. We will present arguments which show that we can expect it to be easier to work with $\mathbf{Y}$ rather than $\mathbf{X}$. To prove positive definiteness we use an approach put forward by Rump in [5]. It relies on a floating point Cholesky decomposition of the midpoint of $\mathbf{Y}$, its backward stability analysis and a perturbation result.

The approach outlined may fail at two places: The variant of Krawczyk’s method may fail to verify the crucial condition (2), or the test for positive definiteness may fail. The chances for failures increase if the condition of $W$ increases. We therefore also present a variant of our algorithm where we use the block diagonalization of Bavely and Stewart [1] to control the condition number of $W$ at the expense of having $D$ with (hopefully small) blocks along the diagonal.

We will present numerical results using Intlab, the Matlab toolbox for interval arithmetic developed by Rump [4]. Since most of our computations rely on matrix-matrix operations, the algorithm is quite efficient. Typically, the computation of the enclosing interval matrix and its test for positive definiteness will not cost more than 5 times the time for computing the (approximate) floating point solution $\tilde{\mathbf{X}}$.

References


Skew-symmetric Matrix Completion for Ranking

David F. Gleich and Lek-Heng Lim

Abstract

We are concerned with the problem of rank aggregation. Given a series of votes on a set of items by a group of voters, rank aggregation is the process of permuting the set of items so that the first element is the best choice in the set, the second element is the next best choice, and so on. This problem is intimately intertwined with the structure of skew-symmetric matrices. Consider a set of \( n \) items, labeled from 1 to \( n \). Suppose that each of these items has an unknown intrinsic quality \( s_i : 1 \leq i \leq n \), where \( s_i > s_j \) implies that item \( i \) is better than item \( j \). While the \( s_i \)'s are unknown, suppose we are given a matrix \( Y \) where \( Y_{ij} = s_i - s_j \). By finding a rank-2 factorization of \( Y \), for example

\[
Y = se^T - es^T,
\]

we can extract unknown scores. The matrix \( Y \) is skew-symmetric and describes any score-based global pairwise ranking. Thus, given a measured \( \hat{Y} \), the goal is to find a minimum rank approximation of \( \hat{Y} \) that models the elements, and ideally one that is rank-2. Phrased in this way, it is a natural candidate for recent developments in the theory of matrix completion.

From this starting point, we make the following contributions. We propose a new method for computing a rank aggregation based on matrix completion, which is tolerant to noise and incomplete data. We solve a structured matrix-completion problem, namely, completion over the space of skew-symmetric matrices:

\[
\begin{align*}
\text{minimize} & \quad \|A(X) - b\|_2 \\
\text{subject to} & \quad \|X\|_* \leq 2 \\
& \quad X = -X^T
\end{align*}
\]

where \( A(\cdot) \) is a linear matrix function that corresponds to only the specified entries of \( \hat{Y} \) and \( \|X\|_* \) is the nuclear norm of a matrix. This problem only differs from the standard matrix completion problem in one regard: the skew-symmetric constraint. With a careful choice of solver, this additional constraint comes “for-free” (with a few caveats). Thus, we extend the SVP algorithm [1] for matrix completion to handle skew-symmetric data and use that to extract ranks for each item. Furthermore, we investigate the results of this algorithm on Netflix movies.

We are currently investigating optimal reconstruction properties of our algorithm and plan to report on those as well.

References

Bounds on Norms of Functions of Matrices Using the Field of Values

Anne Greenbaum

Abstract

Given an $n$ by $n$ matrix $A$ (or a bounded linear operator on a Hilbert space), what can be said about $\|p(A)\|$ where $p$ is a given polynomial or analytic function and $\| \cdot \|$ denotes the operator 2-norm: $\|B\| \equiv \max_{\|v\|=1} \|Bv\|_2$, where $\| \cdot \|_2$ is the 2-norm for vectors.

This question has important applications in a variety of areas. For example, the stability of a linear system of differential equations $y'(t) = Ay(t)$ is governed by the behavior of $\|e^{tA}\|$, and the stability of a finite difference scheme for solving such a system is governed by the norms of powers of the discretization matrix; e.g. $\|(I + (\Delta t)A)^k\|$, $k = 1, 2, \ldots$. Asymptotically (i.e., as $t \to \infty$ or $k \to \infty$), the stability of these systems is determined by the eigenvalues of $A$. If $A$ is normal or near-normal, then it is the eigenvalues that determine the behavior of these matrix functions for all $t$ and $k$, but if $A$ is highly nonnormal then this is not the case.

What other characteristic properties of the matrix (or linear operator) $A$ can be used to bound $\|p(A)\|$? One possibility is the field of values:

$$W(A) = \{q^*Aq : q \in \mathbb{C}^n, \|q\|_2 = 1\}.$$ 

Recently Michel Crouzeix [Bounds for analytical functions of matrices, Integr. Equ. Oper. Theory 48 (2004), pp. 461–477], has made the following conjecture: For any polynomial $p$,

$$\|p(A)\| \leq 2 \max_{z \in W(A)} |p(z)|.$$ 

He was able to prove that the conjecture holds for 2 by 2 matrices and that for a general $n$ by $n$ matrix, $\|p(A)\| \leq 11.08 \max_{z \in W(A)} |p(z)|$. Related results include von Neumann’s inequality (1951): $\|p(A)\| \leq \max_{z \in D(\|A\|, 0)} |p(z)|$, where $D(\|A\|, 0)$ is the disk of radius $\|A\|$ about the origin; the power inequality of Berger and Pearcy (1966): $\|A^k\| \leq 2\nu(A)^k$, where $\nu(A) \equiv \max_{z \in W(A)} |z|$ is the numerical radius; and a result of Okubo and Ando (1975): $\|p(A)\| \leq 2 \max_{z \in D(\nu(A), 0)} |p(z)|$. We discuss these results and attempts to prove or disprove Crouzeix’s conjecture.
Towards Block Structure-Preserving Saddle-Point Solvers

Chen Greif

Abstract

Saddle-point linear systems of the form

\[
\begin{pmatrix}
A & B^T \\
B & 0
\end{pmatrix}
\begin{pmatrix}
u \\
p
\end{pmatrix}
= 
\begin{pmatrix}
f \\
g
\end{pmatrix}
\]

arise in many applications involving problems with constraints. When the matrix is large and sparse, iterative solvers are typically used. A number of Krylov subspace solvers can be applied, but they do not fully take into account the block structure of the matrix. As a result, the underlying Lanczos or Arnoldi process may not give us an easy way to monitor the inertia throughout the iteration. Controlling the inertia is in fact a fairly important goal in many constrained optimization problems. It may therefore be desirable to develop a solver that aims to preserve the saddle-point structure in the approximation subspace. In this talk I will describe a new solver that attempts to accomplish this goal.

As part of the framework, a new block decomposition is derived, which can be used to construct an orthogonal basis for a generalized Krylov subspace associated with mixed powers of matrices related to \( A \) and \( B \). Imposing Galerkin-type conditions, a projection method is derived, which seems to be quite effective in terms of capturing the spectrum of the matrix.

A current weakness of this method is that its performance is not as robust as that of standard Lanczos/Arnoldi-based methods. On the positive side, however, a saddle point structure is maintained in the projected subspace, and matrices such as the Schur complement can be approximated within a modest number of matrix-vector products. For solving linear systems, block preconditioners can be seamlessly incorporated in the iteration.

Numerical experiments for symmetric as well as nonsymmetric saddle point matrices illustrate the convergence properties of the solver and the spectral properties of the matrices that are constructed in the process.
CALU: A Communication Optimal LU Factorization Algorithm

Laura Grigori, James W. Demmel and Hua Xiang

Abstract

Since the cost of communication (moving data) greatly exceeds the cost of doing arithmetic on current and future computing platforms, we are motivated to devise algorithms that communicate as little as possible, even if they do slightly more arithmetic, and as long as they still get the right answer. This talk is about getting the right answer for such an algorithm. It discusses CALU, a communication avoiding LU factorization algorithm based on a new pivoting strategy, that we refer to as tournament pivoting. The reason to consider CALU is that it does an optimal amount of communication, and asymptotically less than Gaussian elimination with partial pivoting (GEPP), and so will be much faster on platforms where communication is expensive, as shown in previous work. The main part of the talk focuses on showing that CALU is stable in practice.

Gaussian elimination with partial pivoting (GEPP) is one of the most stable algorithms for solving a linear system through LU factorization. At each step of the algorithm, the maximum element in each column of $L$ is permuted in diagonal position and used as a pivot. Efficient implementations of this algorithm exist for sequential and parallel machines. In the sequential case, the DGETRF routine in LAPACK implements a block GEPP factorization. The algorithm iterates over block columns (panels). At each step, the LU factorization with partial pivoting of the current panel is computed, a block row of $U$ is determined, and the trailing matrix is updated. Another efficient implementation is recursive GEPP [5, 3]. As discussed in [], we show that DGETRF minimizes neither the bandwidth nor the latency in some cases. Recursive LU attains the bandwidth lower bound but not the latency lower bound in general. In the parallel case, the PDGETRF routine in ScaLAPACK [1] distributes the input matrix over processors using a block cyclic layout. With this partition, every column is distributed over several processors. Finding the maximum element in a column of $L$ necessary for partial pivoting incurs one reduction operation among processors. This gives an overall number of messages at least equal to the number of columns of the matrix. Hence this algorithm cannot attain the latency lower bound of $\Omega(\sqrt{P})$ and is larger by a factor of at least $n/\sqrt{P}$.

CALU uses a new strategy that we refer to as tournament pivoting. This strategy has the property that the communication for computing the panel factorization does not depend on the number of columns. It depends only on the number of blocks in the sequential case and on the number of processors in the parallel case. The panel factorization is performed as follows. A preprocessing step aims at finding at low communication cost $b$ rows that can be used as pivots to factor the entire panel. Then the $b$ rows are permuted into the first positions and the LU factorization with no pivoting of the entire panel is performed. The preprocessing step is performed as a reduction operation with GEPP being the operator used to select pivot rows at each node of the reduction tree. The reduction tree is selected depending on the underlying architecture. In this work we study in particular binary tree based and flat tree based CALU. It has been shown in [2], where the algorithm has been presented for the first time, that binary tree based CALU leads to important speedups in practice over ScaLAPACK on distributed memory computers.

The main part of this work focuses on the stability of CALU. First, we show that the Schur complement obtained after each step of performing CALU on a matrix $A$ is the same as the Schur complement obtained after performing GEPP on a larger matrix whose entries are the same as the entries of $A$ (plus some randomly generated $\epsilon$ entries) and zeros. Hence we expect that CALU will behave as GEPP and it will be also very stable in practice. However, for CALU the upper
bound on the growth factor is worse than for GEPP. The growth factor plays an important role in the backward error analysis of Gaussian elimination. It is computed using the values of the elements of $A$ during the elimination process, \( g_W = \frac{\max_{i,j,k} |a_{ij}^{(k)}|}{\max_{i,j} |a_{ij}|} \), where $a_{ij}$ denotes the absolute value of the element of $A$ at row $i$ and column $j$, and $k$ denotes the matrix obtained at the $k$-th step of elimination. For GEPP the upper bound of the growth factor is $2^{n-1}$, while for CALU it is on the order of $2^{nH}$, where $n$ is the number of columns of the input matrix and $H$ is the depth of the reduction tree. For GEPP the upper bound is attained on a small set of input matrices, that are variations of one particular matrix, the Wilkinson matrix. We show that there are very sparse matrices, formed by Kronecker products involving the Wilkinson matrix, that nearly attain the bound. Moreover, there are Wilkinson-like matrices for which GEPP is stable and CALU has exponential growth factor and vice-versa.

Second, we present experimental results for random matrices and for a set of special matrices, including sparse matrices, for binary tree based and flat tree based CALU. We discuss both the stability of the LU factorization and of the linear solver, in terms of pivot growth and backward errors. The results show that in practice CALU is stable. They show that CALU leads to backward errors within a factor of 10 of the GEPP backward errors (except for one matrix for which the ratio of the normwise backward error of CALU to GEPP is within a factor of 26).

We also discuss the stability of block versions of pairwise pivoting [4] and parallel pivoting [6], two different pivoting schemes. These methods are of interest, since with an optimal layout, block pairwise pivoting is communication optimal in a sequential environment and block parallel pivoting is communication optimal in a parallel environment. It is simple to see that block parallel pivoting is unstable. With an increasing number of blocks per panel (determined by the number of processors), the growth factor is getting larger. In the extreme case when the block size is equal to 1, the growth factor is exponential on random examples. For pairwise pivoting we study the growth factor for the case when the block size is equal to 1. This method is more stable, but it shows a growth more than linear of the factor with respect to the matrix size. Hence a more thorough analysis for larger matrices is necessary to understand the stability of pairwise pivoting.

References

Basic Iterative Algorithms of Numerical Linear Algebra as Building Blocks for Hybrid Adaptive Finite Element Methods

Randolph E. Bank, Luka Grubišić, Agnieszka Miedlar and Jeffrey S. Ovall

Abstract

Implementation and analysis of modern adaptive finite element procedures for the computation of eigenvalues of partial differential operators requires a fine interplay of techniques from approximation theory, analysis, linear algebra and discrete mathematics. Often computational effort spent on doing linear algebra tasks far outweighs the cumulative effort in performing other algorithmic steps. Yet, utilization of linear algebra is seldom optimized as an integral part the adaptive mesh refinement loop Compute $\rightarrow$ Estimate $\rightarrow$ Mark $\rightarrow$ Refine. We will call the standard approach to adaptive refinement under the assumption of exact linear algebra AFEM. In contrast, we distinguish an adaptive finite element procedure with a balanced, possibly inexact, linear algebra part as AFEMLA approach. For further discussion on this topic see [5].

Our approach to the problem of designing efficient AFEMLA algorithms is the following. First, we present a general framework for a posteriori estimation of the error and convergence enhancement in eigenvalue/eigenvector computations for symmetric and elliptic eigenvalue problems, and provide detailed analysis of a specific and important example within this framework — finite element methods with continuous affine elements, see references [2, 3]. A distinguishing feature of the proposed approach is that it provides provably efficient and reliable error estimation under very realistic assumptions, not only for single, simple eigenvalues, but also for clusters which may contain degenerate eigenvalues. We reduce the study of the eigenvalue/eigenvector error estimators to the study of associated boundary value problems, and make use of the wealth of knowledge available for such problems. Our choice of a posteriori error estimator, computed using hierarchical bases, very naturally offers means not only for estimating error in eigenvalue/eigenvector computations, but also cheaply accelerating the convergence of these computations — sometimes with convergence rates which are nearly twice that of un-accelerated approximations.

Second, we show that in fact our estimator has been computed by an implicit application of several basic iterative algorithms of numerical linear algebra. In particular the preconditioned inverse iteration features prominently both in the context of the computation of mesh refinement indicators as well as in the context of convergence enhancement. Further favorable properties of inverse iteration in AFEM context have recently been analyzed in [7].

Building upon the simple algebraic structure of our hybrid algorithm we are able to include elliptic non-self adjoint problems — using the technique of [4] — as well as inexact adaptive eigenvalue approximation techniques like AFEMLA algorithm in our theoretical framework. Naturally, this hierarchical extension incurs extra costs when compared to running only standard sparse eigenvalue solvers. However, we argue that extra effort invested in computing a residual representation, even in the context of inexact adaptive eigenvalue/vector solver, can effectively be traded off for increased convergence performance of the accelerated Ritz values/vectors. In comparison a competing Ritz value acceleration and estimation technique can be designed based on the gradient recovery algorithms from e.g. [6]. However, the gradient recovery algorithms — as the name says — recover gradients and so they cannot be readily applied to construct Ritz vector enhancement. We construct Ritz vector enhancements as a by product when using the hierarchical bases technique from [1]. We further reuse this enhanced Ritz vector as an ever improving starting vector for the eigenvalue iterative algorithm on the next mesh refinement level. This further improves the overall efficiency of the algorithm. We report numerical experiments in detail.
References


Reduced Rank Regression via Convex Optimization

Ming Gu

Abstract

We consider the classical multivariate least squares regression problem:

\[ \min_X \|AX - B\|_F, \]  

(1)

where \( A \in \mathbb{R}^{n \times m} \) and \( B \in \mathbb{R}^{n \times l} \) are given matrices, and \( X \in \mathbb{R}^{m \times l} \) is the unknown matrix. In this abstract, we assume \( n > m > l \).

To solve equation (1), let \( A = QR \) be the QR factorization of \( A \) with \( Q \in \mathbb{R}^{n \times m} \) and \( R \in \mathbb{R}^{m \times m} \). Then the solution \( X = R^{-1} (Q' B) \).

While such a solution is relatively straightforward to compute, one potential flaw with it is that in the case where \( l \) is relatively large, problem (1) may be an over-parameterization, and its solution may be too noisy to be useful in practice. As a consequence, we would like to solve a restricted version of (1) in the following form:

\[ \min_X \|AX - B\|_F, \quad s.t. \quad \text{rank}(X) \leq r, \]  

(2)

where \( r \) is a given positive integer. Problem (2) is known as reduced rank regression (RRR) in statistics, and is widely used in practice. There are a number of good statistical reasons why this reduced rank model is better. In addition, the reduce rank regression model solution can be efficiently computed. Let \( QR = A \) be the QR factorization of \( A \) as before, and let the rank \( r \) truncated SVD of the matrix \( Q'B \) be \( U_r S_r V_r^T \). Then the solution is simply \( X = (R^{-1} U_r) (V_r S_r)^T \).

The parentetheses in this expression are used to reduce the overall computation.

Recently, with the rapid progress of efficient algorithms and interesting applications for comressive sensing ([1]), it is proposed ([2]) that one can study a convex optimization alternative to the classical reduced rank regression problem (2):

\[ \min_X \|AX - B\|_F, \quad s.t. \quad \|X\|_* \leq \tau, \]  

(3)

where \( \tau > 0 \) is a given parameter, and \( \|X\|_* \) is the so called \textit{nuclear norm}, which is the sum of all the singular values of \( X \).

The motivation behind (3) is two-fold:

- While the classical RRR model solution depends discontinuously on the rank \( r \), the new model solution depends smoothly on the parameter \( \tau \). This can be very important in situations where the parameter estimation problem is adaptively solved, such as in the case of time series analysis.

- Efficient algorithms are known to exist for such convex programs.
There are several equivalent formulations of the new RRR model ([2, 3]). Similar to work in compressed sensing, we call the new RRR the matrix Lasso problem or the matrix $LS_\tau$ problem. Another way to formulate the RRR problem is
\[
\min_X \frac{1}{2} \|AX - B\|_F^2 + \lambda \|X\|_1, \tag{4}
\]
for a given value of $\lambda$. This is called a matrix QP$_\lambda$ problem. It is known that for any given $\lambda > 0$, there exists a $\tau$ so that both matrix $LS_\tau$ and matrix QP$_\lambda$ problems have exactly the same solution $X$.

Yet a third formulation is the following minimization problem:
\[
\min_X \|X\|_1, \tag{5}
\]
\[
\text{s.t. } \|AX - B\|_F \leq \sigma,
\]
where $\sigma > 0$ is a given parameter. This is called a matrix BP$_\sigma$ (Basis Pursuit denoise) problem. Again for any given $\sigma$ there exists a $\tau > 0$ so that both the matrix BP$_\sigma$ problem and the matrix $LS_\tau$ problem have the same solution.

Yuan, et, al developed an algorithm based on the matrix QP$_\lambda$ problem ([2]). They show that their algorithm is much more efficient than the SDP3 algorithm for solving standard SDP (Semi-Definite Programming) problems, both in execution time and memory requirements.

The algorithm spgl1, developed by Berg and Friedlander ([4]), is a well-known algorithm for solving the vector BP$_\sigma$ problem
\[
\min_x \|x\|_1, \tag{6}
\]
\[
\text{s.t. } \|Ax - b\|_F \leq \sigma,
\]
where $b$ is a given vector and $x$ a solution vector. We show that spgl1 can be trivially generalized to solve the matrix BP$_\sigma$ problem with equally impressive performance. In particular, for each of the test problems used in ([3]), our generalization of spgl1 takes less than 1 second in matlab to solve the matrix BP$_\sigma$ version, whereas the algorithm in ([3]) takes up to several hundred seconds in C to solve the matrix QP$_\lambda$ version.

References:


Towards Black-Box Rational Krylov Methods for $f(A)b$
Automated Parameter Selection for Markov Functions and Error Estimation

Stefan Güttel

Abstract

Given a square matrix $A$, a vector $b$, and a function $f(z)$ analytic on the eigenvalues $\Lambda(A)$, we are interested in the efficient computation of the vector $f(A)b$. Of particular importance in scientific computing are the functions $f(z) = (z - i\omega)^{-1}$ or $f(z) = \exp(tz)$ (dynamical systems in the frequency or time domain), or $f(z) = z^{-1/2}$ (Dirichlet-to-Neumann maps) [9]. If $A$ is large and sparse then $f(A)$ is generally a full matrix, the computation and storage of which are prohibitive. A popular approach for efficiently extracting an approximation from a rational Krylov space [12, 13] is the rational Arnoldi method [4, 3, 8, 6] (or variants [11, 7] thereof).

Three steps. Here is a three-step description of the rational Arnoldi method:

1. Select a sequence $\{\xi_1, \xi_2, \ldots\}$ of complex numbers (the “poles”), disjoint with $\Lambda(A)$.

2. Compute an orthonormal basis $V_n = [v_1, \ldots, v_n]$ of a rational Krylov space: Set $v_1 = b/\|b\|$, then orthonormalize $w_{j+1} = (A - \xi_j I)^{-1}v_j$ against $\{v_1, \ldots, v_j\}$, which gives $v_{j+1}$.

3. Compute the $n$-th order Arnoldi approximation for $f(A)b$ as $f_n := V_n f(A_n) V_n^H b$, $A_n := V_n^H A V_n$,

which only requires the evaluation of a small $n \times n$ matrix function.

It is known that (in exact arithmetic) this method implicitly computes a quasi-optimal approximation $f_n = r_n(A)b$ for $f(A)b$, where $r_n(z)$ is a rational function with prescribed poles $\xi_j$ [4, 3, 8].

Three problems. Each step of the rational Arnoldi method above raises at least one problem, each of which prevents this method from being used more widely in practice. We intend to answer the following questions on our poster and thereby propose a “black-box” rational Krylov method.

1. How to select the pole sequence $\{\xi_1, \xi_2, \ldots\}$?

We propose a heuristic adaptive pole selection strategy for Markov functions $f(z)$ (such as $f(z) = z^{-1/2}$), inspired by ideas proposed for the exponential [5] and transfer functions [6]. In contrast to the situation for these special functions, the condenser on which our underlying rational approximation process takes place is not symmetric with respect to the imaginary axis. Instead, one condenser plate is the support of the generating measure for $f(z)$ (in our case the negative real axis), and the other plate is the numerical range of $A$. Since the numerical range of $A$ is not known a-priori, we “discretize” it by rational Ritz values of one order higher than the current iteration index. This results in a new efficient variant of the rational Arnoldi method which is completely free of parameters. Numerical experiments indicate that convergence is at least as good as if we were using generalized Leja poles on the continuous condenser. Recent results on the convergence of rational Ritz values give some insight on how our method may even produce superlinear convergence [1, 2].
2. What happens if the linear systems for $w_{j+1}$ are solved inexactly?

An inexact Arnoldi approximation $\tilde{f}_n$ can be interpreted as an exact Arnoldi approximation for a nearby problem $f(A+D_n)b$, where $D_n$ is some correction matrix of small norm (cf. [10]).

Knowing the residuals of the linear systems solved in the rational Arnoldi method allows us to easily compute a corrected Arnoldi approximation $\hat{f}_n$, which is an approximation obtained by orthogonal reduction of $f(A)b$ with respect to an inexact search space. The distance $\|\hat{f}_n - \tilde{f}_n\|$ is a practical estimator for the error introduced by inexact linear system solves. We refer to this error as sensitivity error, because the norm $\|f(A)b - f(A + D_n)b\|$ depends on the sensitivity of $f(z)$ with respect to perturbations in the argument. For more details see [8].

3. How to estimate the error $\|f(A)b - f_n\|$? (There is no simple residual equation!)

The error $\|f(A)b - f_n\|$ consists of two parts

$$\|f(A)b - f_n\| \leq \|f(A)b - f(A + D_n)b\| + \|f(A + D_n)b - \tilde{f}_n\|.$$  

Once the sensitivity error is estimated (see 2.), the approximation error can be estimated by techniques known for rational interpolants with prescribed poles. Of course, there is no reason to continue iterating once the approximation error has fallen below the sensitivity error. This yields a practical stopping criterion for the rational Arnoldi method.

References


Optimal $\mathcal{H}_2$ Points in Action: From Model Reduction to Lebesgue Constant

Christopehr A. Beattie, Garret Flagg, and Serkan Gugercin

Abstract

Consider a single-input/single-output (SISO) linear dynamical system

$$
G : \begin{cases}
\dot{x}(t) = Ax(t) + bu(t) \\
y(t) = c^T x(t)
\end{cases}
$$

where $A \in \mathbb{R}^{n \times n}$, $b, c \in \mathbb{R}^n$; $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}$, $y(t) \in \mathbb{R}$, are the state, input, and output, respectively, of the system. Extensions to multi-input/multi-output case are possible; however for the brevity of this abstract, we will focus on the SISO case. In most applications, $n$ is quite large, resulting in untenable demands on computational resources. One remedy is model reduction. The goal is to construct a reduced-order model satisfying the first-order optimality conditions $[5]$. IRKA has been developed to generate a reduced-order model satisfying the first-order $\mathcal{H}_2$ optimality conditions $[5]$. IRKA has been applied in settings with very large state-space dimensions efficiently producing high-fidelity reduced-order models. We will call the interpolation points satisfying the $\mathcal{H}_2$-optimal conditions as $\text{IRKA-points}$.

In this talk, we show that IRKA-points prove useful in many other applications different than their initial $\mathcal{H}_2$-optimal setting. The first such application is approximation in the $\mathcal{H}_\infty$ norm, the $2-2$
induced norm of the convolution operator, i.e. the uniform norm. We will show that IRKA-points provide a very effective initialization strategy for an interpolation-based approach to $H_\infty$ norm approximation problem. The resulting reduced-order model is an interpolatory high-fidelity $H_\infty$ approximant, leading to a near-circular error curve behavior as desired [7].

The second application arises in analyzing the perturbation effects of iterative solves (as opposed to direct solves) in interpolatory model reduction. The norm of the underlying oblique projector in going from $G(s)$ to $G_r(s)$ plays a vital role in this setting [2]. The larger the projector norm, the bigger the perturbation error, hence the worse the approximation. While this norm grows extremely rapidly for arbitrarily chosen interpolation points, it stays impressively modest for IRKA-points. This immediately reminds the concept of Lebesgue constant; where the constant grows exponentially for equidistant nodes but grows only logarithmically for the Chebyshev nodes; e.g. see [3]. We will point out the similarities/connections between these two settings.

Finally, we will look at the IRKA-points in the ADI setting for solving Lyapunov equations. We prove that once the ADI shifts are chosen using IRKA-points, the multi-shift ADI step is equivalent to solving a projected Lyapunov equations and then pulling it back up to the original dimension. We illustrate examples where IRKA-points based ADI shifts shows fast convergence behavior. We note that since the IRKA-points are obtained by solving a sequence of linear systems, we are not necessarily proposing a numerical methodology but rather illustrating a theoretical concept. Then using the ADI minimax problem behind the ADI-shift selection, we illustrate the potential-theoretic properties of the IRKA-points and their connection to the rational Zolotarev problem.

References


On a Nonlinear Matrix Equation Arising in Nano Research

Chun-Hua Guo, Yueh-Cheng Kuo, and Wen-Wei Lin

Abstract

The matrix equation $X + A^T X^{-1} A = Q$ arises in Green’s function calculations in nano research, where $A$ is a real $n \times n$ matrix and $Q$ is a real symmetric matrix dependent on a parameter and is usually indefinite. In practice one is mainly interested in those values of the parameter for which the matrix equation has no stabilizing solutions. The solution of interest in this case is a special weakly stabilizing complex symmetric solution $X^*$, which is the limit of the unique stabilizing solution $X_\eta$ of the perturbed equation $X + A^T X^{-1} A = Q + i\eta I$, as $\eta \to 0^+$. It has been shown in [1] that a doubling algorithm can be used to compute $X_\eta$ efficiently even for very small values of $\eta$, thus providing good approximations to $X^*$. It has been observed by nano scientists that a modified fixed-point method can sometimes be very efficient, particularly for computing $X_\eta$ for many different values of the parameter. We provide a rigorous analysis of this modified fixed-point method and its variant, and of their generalizations. We also show that the imaginary part $X_I$ of the matrix $X^*$ is positive semidefinite and determine the rank of $X_I$ in terms of the number of unimodular eigenvalues of the quadratic pencil $P(\lambda) = \lambda^2 A^T - \lambda Q + A$. Finally we present a structure-preserving algorithm that can find $X^*$ directly from the equation $X + A^T X^{-1} A = Q$, under the assumption that all unimodular eigenvalues $\lambda \neq \pm 1$ of $P(\lambda)$ are semisimple and the eigenvalues $\pm 1$ (if any) have partial multiplicities 2. The algorithm computes $X^*$ by finding a suitable invariant subspace of a symplectic pencil $\mathcal{M} - \lambda \mathcal{L}$ that is a linearization of $P(\lambda)$. It starts with using the $(S + S^{-1})$-transform in [2] to transform $\mathcal{M} - \lambda \mathcal{L}$ into a new pencil $\mathcal{K} - \lambda \mathcal{N}$, where $\mathcal{K}$ and $\mathcal{N}$ are both skew-Hamiltonian. This new pencil is then reduced using the Patel approach in [3]. A suitable invariant subspace of $\mathcal{K} - \lambda \mathcal{N}$ is then obtained by performing eigen computations for an $n \times n$ matrix pencil, and is converted to the required invariant subspace of the pencil $\mathcal{M} - \lambda \mathcal{L}$. The computational work of this algorithm is roughly 1/4 of that for the QZ algorithm applied directly on the $2n \times 2n$ pencil $\mathcal{M} - \lambda \mathcal{L}$. Numerical results show that the accuracy of $X^*$ from the structure-preserving algorithm is as good as that from the QZ algorithm.

References


A Fast Algorithm for Approximating the Distance to Instability

Mert Gurbuzbalaban and Michael L. Overton

Abstract

An \( n \times n \) complex matrix \( A \) is stable if and only if the spectral abscissa of \( A \) (the maximum of the real parts of the eigenvalues, denoted \( \alpha \)), is negative. However, \( \alpha \) is not a robust measure of stability, as small perturbations to the matrix can result in large changes to the eigenvalues. A natural alternative measure is the distance to (continuous) instability of \( A \), that is, the distance from \( A \) to the set of unstable matrices:

\[
d_{CI}(A) = \sup \{ \epsilon : A + E \text{ is stable for all } E \text{ with } \|E\| \leq \epsilon \}
\]

where \( \| \cdot \| \) denotes the matrix 2-norm. In the systems and control theory community, \( d_{CI} \) is known as the complex stability radius \([4]\). Byers [1] gave the first algorithm for computing \( d_{CI} \) to arbitrary accuracy; this bisection method introduced the key idea of determining whether or not \( d_{CI}(A) \geq \delta \) for a given \( \delta \) by verifying whether any of the eigenvalues of a certain Hamiltonian matrix are imaginary. Soon afterwards, quadratically convergent variations appeared; these were developed for the more general problem of computing the \( H_{\infty} \) norm of the transfer function for a linear dynamical system with input control and output measurement. These algorithms depend on computing eigenvalue decompositions of Hamiltonian matrices of order \( 2n \) and SVDs of order \( n \), so the cost is effectively \( O(n^3) \) and they are applicable only to small or moderate-sized systems.

In order to handle larger matrices, we propose a fast algorithm that avoids Hamiltonian eigenvalue decompositions and SVDs. The only allowable matrix operation is computing the rightmost eigenvalue \( \lambda \) of \( X \) and its corresponding right and left eigenvectors, which is done by an iterative algorithm such as implicitly restarted Arnoldi or biorthogonal Lanczos using matrix-vector products. (A rightmost eigenvalue \( \lambda \) of \( X \) is one for which \( \text{Re} \lambda = \alpha(X) \).) Our method is inspired by the following recent algorithm [2] developed to compute the \( \epsilon \)-pseudospectral abscissa, denoted \( \alpha_\epsilon \). The quantity \( \alpha_\epsilon(A) \) is the largest of the real parts of the elements in the \( \epsilon \)-pseudospectrum \( \Lambda_\epsilon(A) \) (the set of points in the complex plane which are eigenvalues of \( A + \Delta A \) for some \( \Delta A \) with \( \| \Delta A \| \leq \epsilon \)).

Algorithm 1

1. Let \( z_0 \) be an eigenvalue of \( A \), with corresponding right and left eigenvectors \( x_0 \) and \( y_0 \) normalized so \( \|x_0\| = \|y_0\| = 1 \) and \( y_0^* x_0 \in (0,1) \). If \( \alpha(A + \epsilon_0 \ y_0^* x_0^* ) \geq \text{Re}(z_0) + \epsilon_0 y_0^* x_0 \), set \( E_1 = y_0^* x_0^* \); otherwise, set \( E_1 = x_0 y_0^* \). Set \( B_1 = A + \epsilon_0 E_1 \).

2. For \( k = 1, 2, \ldots \), let \( z_k \) be the rightmost eigenvalue of \( B_k \), with right and left eigenvectors \( x_k \) and \( y_k \) normalized so that \( \|x_k\| = \|y_k\| = 1 \) and \( y_k^* x_k \in (0,1) \). Set \( E_{k+1} = y_k^* x_k^* \) and \( B_{k+1} = A + \epsilon_k E_{k+1} \). If \( |\alpha(B_{k+1}) - \alpha(B_k)| \leq \tau \), stop.

Algorithm 1 is guaranteed to generate a sequence of lower bounds for \( \alpha_\epsilon(A) \). A detailed analysis of sufficient conditions for these lower bounds to converge to \( \alpha_\epsilon(A) \) is given in [2].

One approach to computing \( d_{CI}(A) \) would be to repeatedly use Algorithm 1 to compute \( \alpha_{\epsilon_j}(A) \) for a sequence \( \epsilon_j \), using bisection on \( \epsilon \), but this would introduce a third nested iteration and also requires the assumption that each \( \alpha_{\epsilon_j}(A) \) is computed accurately. Instead, we wish to compute \( d_{CI} \) more directly by generating a sequence \( B_k = A + \epsilon_k E_k \) with \( \epsilon_k \) converging to \( d_{CI}(A) \). This necessitates varying \( \epsilon_k \) at each step \( k \) in a subtle way. We set \( \epsilon_{k+1} = \epsilon_k - (y_k^* x_k^*) \alpha(B_k) \) and motivate this choice by a local analysis near the optimizer under some regularity assumptions on the abscissa. This leads to the following algorithm:
Algorithm 2

1. Let \( z_0 \) be an eigenvalue of \( A \), with corresponding right and left eigenvectors \( x_0 \) and \( y_0 \) normalized so that \( \|x_0\| = \|y_0\| = 1 \) and \( y_0^* x_0 \in (0, 1) \). If \( \alpha(A + \epsilon y_0 x_0^*) \geq \text{Re}(z_0) + \epsilon \text{Re}(y_0^* x_0) \), set \( E_1 = y_0^* x_0 \); otherwise, set \( E_1 = x_0 y_0^* \). Set \( B_1 = A + \epsilon E_1 \) with \( \epsilon = -(y_0^* x_0) \text{Re}(z_0) \).

2. For \( k = 1, 2, \ldots \) let \( z_k \) be the rightmost eigenvalue of \( B_k \), with right and left eigenvectors \( x_k \) and \( y_k \) normalized so that \( \|x_k\| = \|y_k\| = 1 \) and \( y_k^* x_k \in (0, 1) \). Set \( E_{k+1} = y_k^* x_k \), \( \epsilon_{k+1} = \epsilon_k - (y_k^* x_k) \alpha(B_k) \), and \( B_{k+1} = A + \epsilon_{k+1} E_{k+1} \). If \( |\epsilon_{k+1} - \epsilon_k| \leq \tau \), stop.

We will discuss the convergence of the algorithm and prove conditions under which \( \epsilon_k \) converges to \( d_{CI}(A) \). See the website www.cims.nyu.edu/~mert/householder.pdf for figures that illustrate the convergence behavior. Very often, \( \epsilon_k \) converges to \( d_{CI}(A) \). In some cases Algorithm 2 may overestimate the stability radius; this happens when \( z_k \) converges to a local maximizer of the pseudospectral-abscissa (as can happen with Algorithm 1 when \( \epsilon \) is not sufficiently small). In these cases, \( d_{CI}(A) \) will be poorly approximated. However, global optimality can be checked a posteriori by computing a single \( 2n \times 2n \) Hamiltonian eigenvalue decomposition, and if necessary the algorithm can be restarted from another eigenvalue. Of course, this is not possible if \( n \) is sufficiently large, but in that case the standard algorithms are inapplicable too. Ironically, Algorithm 2 might turn out to be more reliable than some standard implementations of the algorithms that compute eigenvalues of Hamiltonian systems which are also prone to underestimating \( d_{CI}(A) \); this is because lower bounds for \( d_{CI} \) are obtained by verifying that a Hamiltonian matrix has no imaginary eigenvalues with a certain property, and if the methods that compute the eigenvalues of the Hamiltonian matrix are not guaranteed to compute simple imaginary eigenvalues with zero real rounding errors, then an eigensolver may erroneously conclude that a matrix does not have an imaginary eigenvalue when a more careful implementation would have determined that it does.

In the definition of \( d_{CI} \), complex perturbations are admitted. For a real matrix \( A \), if we only permit real perturbations, the resulting stability measure \( d_{CI}^R \) is well known to be more difficult to compute. We discuss how to approximate \( d_{CI}^R \) by a simple variation of Algorithm 2. We also discuss extension of Algorithm 2 to the equally important problem of computing \( d_{DI}(A) = \sup \{ \epsilon : |z| < 1 \text{ for all } z \in \Lambda_\epsilon(A) \} \) for which the boundary of the stability region is the unit circle instead of the imaginary axis. Finally, our longer term goal is to develop a fast algorithm for the computation of the \( H_\infty \) norm for large dynamical systems, which could open the way to extending software such as HIFOO [3] to much larger systems such as those that are governed by partial differential equations.

References


Spectral Deflation in Krylov Solvers

André Gaul, Martin H. Gutknecht, Jörg Liesen, and Reinhard Nabben

Abstract

Krylov solvers for linear systems of equations tend to converge slowly — or not at all — if the system matrix \( A \) has eigenvalues of small absolute value, a situation that is very common in practice. The standard remedy is to apply preconditioning of some sort. A particularly effective technique for dealing with such problems is to identify an approximate eigenspace \( Z \) that belongs to a set of such small eigenvalues and to apply spectral deflation to \( A \) combined with an augmentation of the bases of the Krylov subspaces. By using an orthogonal projection \( P \) with nullspace \( Z \), and by restricting the operator \( A \) accordingly, the Krylov solver can be applied in the orthogonal complement \( Z^⊥ \) only. The bases constructed implicitly or explicitly by this restricted operator are augmented by a set of basis vectors for \( Z \). Various ways to handle and implement this general approach have been proposed, starting with the deflated (and augmented) three-term CG version of Nicolaides ’85/’87 SINUM (the first number indicates the year the paper was submitted). Several authors (e.g., Morgan ’93/’95 SIMAX, de Sturler ’93/’96 JCAM, and Chapman/Saad ’95/’97 NLAA) described versions of GMRES with augmented basis but no deflation of the operator. Others, such as Kharchenko/Yeremin ’92/’95 NLAA and Erhel/Burrage/Pohl ’94/’96 JCAM proposed a GMRES where the small eigenvalues are shifted by a right-hand side preconditioner to a large positive value. All these GMRES versions differ not only algorithmically and numerically, but often also mathematically, in particular when \( Z \) is not exactly \( A \)-invariant. Many more contributions to this topic have been made in the last fifteen years. Some keywords associated with such methods are ‘(spectral) deflation’, ‘augmented basis’, ‘recycling Krylov subspaces’, and ‘singular preconditioning’. In fact, the deflated operator \( PA \) can be viewed as left-side preconditioned operator with a singular preconditioner.

More recently, in a series of papers by varying subsets of a group comprising Erlangga, MacLachlan, Nabben, Vuik, and Tang, it was shown that from an algebraic viewpoint, two-level methods based on deflation, domain decomposition, or multigrid are of the same structure. In this talk we will concentrate on Krylov solvers that incorporate both augmentation of the bases and deflation of the operator. In contrast to the previous work mentioned above it will include (if time permits) at once

1. a clean derivation of two different approaches: one based on orthogonal projections (most appropriate for Hermitian operators) and one based on oblique projections (adapted to non-Hermitian operators),

2. a set of analogous descriptions of various coordinate space based methods: deflated GMRES and MINRES, truly deflated GMRES, and deflated QMR,

3. deflated CG (old) and deflated BiCG (new), as well as other deflated methods based on the nonsymmetric Lanczos process, in particular Lanczos-type product methods such as BiCGStab.

Our main messages is that, from a theoretical point of view, if \( A \) is non-Hermitian, \( P \) should be chosen as a suitable oblique projection, so that when the nullspace \( \tilde{Z} \) of \( \tilde{A} := PAP \) is \( A \)-invariant (in a particularly strong sense in case of multiple eigenvalues), so is its range \( \tilde{Z}^⊥ \), and the restrictions of \( A \) and \( \tilde{A} \) to \( \tilde{Z}^⊥ \) are identical:
(i) \( Z \mathbf{A}-\text{invariant} \iff \tilde{Z}^\perp \mathbf{A}-\text{invariant} \),
(ii) \( Z \mathbf{A}-\text{invariant} \implies \tilde{\mathbf{A}}|_Z = \mathbf{O}|_Z, \quad \tilde{\mathbf{A}}|_{\tilde{Z}^\perp} = \mathbf{A}|_{\tilde{Z}^\perp} \).

This is a true benefit of the approach based on oblique projections. When using orthogonal projections, we can only match the eigenvalues of the restrictions to \( Z^\perp \), not the corresponding eigenvectors and principal vectors. That is, in general, \( \mathbf{A} \) and \( \tilde{\mathbf{A}} \) differ on \( Z^\perp \).

So far, this point got little emphasis in the literature, except for a short QCD conference proceedings paper by Abdel-Rehim, Morgan, and Wilcox (\textit{Lattice/07}), where a deflated BiCGStab has been sketched, but the above described rationale has not been mentioned.
Spectral Relaxations of Hard Combinatorial Problems

Christopher J. Hillar, Kilian Koepsell, and Friedrich Sommer

Abstract

In many domains, it is useful to isolate the interacting components of a given process, usually by analyzing data generated from it. One mathematical formulation of the problem uses the framework of graph clustering. We shall outline some powerful new spectral techniques that can solve clustering problems in practice, and then explore many interesting mathematical questions about matrices that arise in this context.

Let $G$ be a weighted graph on vertices $V = \{1, \ldots, n\}$ with edge weights $a_{ij} \geq 0$ encoded by a symmetric similarity matrix $A = A^\top \in \mathbb{R}^{n \times n}$. The numbers $a_{ij}$ indicate the strength of the relationship or similarity between vertices $i$ and $j$. In very general terms, the goal of network (or community) clustering is to find subsets of the vertices which have high similarity. For example, in the setting of neuroscience, vertices represent retinal ganglion cells and the edge weights correspond to (data-inferred) coupling strengths between neighboring neurons. The goal is then to find highly connected subsets of interacting neurons, called cell assemblies.

When $G$ is a normal graph and $A$ is its adjacency matrix, the problem becomes that of finding large cliques. The computational complexity of determining the largest clique size (called the MAXCLIQUE problem) in a graph is known to be NP-complete [4] and NP-hard to approximate [3, 8]. Thus, unless $P = NP$, there is likely no fast method to compute the largest clusters in general graphs. However, in recent years, methods from spectral graph theory have been finding large clusters experimentally, even for big graphs. For instance, in the context of image segmentation, Shi and Malik [7] (see also the work of [1]) use the eigenvectors of normalized Laplacians to find large segments in images. Here, we shall focus on the very recent modularity method which has been used successfully to detect community structure in networks [5] (see also [6] and the references therein).

Let $G$ be an ensemble of weighted graphs (i.e. a probability distribution over $n \times n$ similarity matrices $W$) and let $G$ be a graph drawn from $G$. Given any two vertices $i$ and $j$ of $G$, we may compare their edge strength $a_{ij}$ to the expected edge weight $E[w_{ij}]$ between $i$ and $j$ over the entire ensemble $G$. The difference

$$M_{ij} := a_{ij} - E[w_{ij}]$$

gives a measure of surprise of the weight $a_{ij}$ given the ensemble $G$. Given a potential partitioning $\{C, \overline{C}\}$ of $V$ into a cluster $C \subseteq V$ and its complement $\overline{C}$, the modularity of $C$ is

$$M(C) := \sum_{i,j \in C} M_{ij} + \sum_{i,j \in \overline{C}} M_{ij}. \quad (1)$$

The quantity $M(C)$ is large when there are edge strengths between nodes in both communities $C$ and $\overline{C}$ that are much larger than expected. Although maximizing modularity is known to be NP-complete in its decision formulation [2], clusters with experimentally large modularity can be found using a spectral relaxation. The idea is to first encode the cluster $C$ with a bipolar column vector $c = [c_1, \ldots, c_n]^\top \in \{-1, 1\}^n$ (the positive support of $c$ defines the partition) and then rewrite the modularity as a quadratic form:
\[ M(C) = \sum_{i,j=1}^{n} M_{ij}(1 + c_i c_j) = \frac{1}{2} c^\top M c + \text{const.} \] (2)

Next, we relax the problem by maximizing the quantity \( M(C) \) over real vectors \( c \in \mathbb{R}^n \) instead of the finite combinatorial set of partitions of \( V \). The cluster is then determined by taking the sign of \( c \) (i.e. replace positive components of \( c \) with 1’s and negatives ones with \(-1\)’s). In symbols,

\[
\text{arg max}_{C \subseteq V} M(C) = \text{arg max}_{c \in \{-1,1\}^n, c^\top c = n} c^\top M c \approx \text{sgn} \text{ arg max}_{c \in \mathbb{R}^n, c^\top c = n} c^\top M c = \text{sgn} \text{ arg max}_{0 \neq c \in \mathbb{R}^n} \frac{c^\top M c}{c^\top c}. \] (3)

The last expression above equals the sign of the eigenvector corresponding to this largest eigenvalue \( \lambda \). Seemingly crude, this relaxation (to optimize a Rayleigh quotient) produces subsets of vertices with high modularity. In fact, it can be shown that the approximation (i) is independent of the labeling of vertices, (ii) is exact for many natural graphs, and (iii) produces intuitively correct clusters in these exact cases. We explain the precise nature of these results and pose many natural spectral problems in this setting. This exploration answers a call in [6] for a theoretical explanation of the experimental findings using these methods.

References


The Golub-Kahan Iterative Bidiagonalization in Regularization of Ill-posed Problems and Estimation of the Noise in the Data

Iveta Hnětynková, Martin Plešinger, and Zdeněk Strakoš

Abstract

A broad class of applications requires solving linear ill-posed approximation problems of the form

$$Ax \approx b$$

with a right-hand side $b$ (observation vector) contaminated by noise. The matrix $A$ often represents a discretized smoothing operator (such as a discretized blurring operator in image deblurring problems) with the singular values of $A$ decaying gradually without a noticeable gap; $A$ is usually ill-conditioned. The presence of the noise causes additional difficulties; the direct solution $A^+x$, where $A^+$ denotes a matrix pseudoinverse, represents usually a meaningless noise-dominated solution. Therefore regularization methods are used for finding numerical approximations to the solution which reflect a sufficient amount of information contained in the data, while suppressing the devastating influence of the noise.

The Golub-Kahan iterative bidiagonalization belongs among popular techniques with regularization properties. Here the original problem is projected onto a lower dimensional (Krylov) subspace, which in fact represents a form of regularization by projection. The projected problem, however, inherits a part of the ill-posedness of the original problem, and therefore some inner regularization must be applied [5]. Stopping criteria for the whole process are usually based on the regularization of the projected (small) problem. Regularization parameters are typically determined by L-curve techniques, estimation of the distance between the exact and regularized solution, the discrepancy principle, cross validation methods (see, e.g., [1, Chap. 7, pp. 175–208] and [6] for comparison of these methods).

In this contribution we restrict ourselves to ill-posed problems where the right-hand side $b$ is contaminated by white noise

$$b = b^{\text{exact}} + b^{\text{noise}}$$

with the unknown noise level; we only assume $\|b^{\text{noise}}\| \ll \|b^{\text{exact}}\|$. By the nature of the problem we can assume that multiplication of a vector $v$ by $A$ and $A^T$ results in smoothing which reduces the relative sizes of the high frequency components in $v$. In addition we assume that (on average) the left singular vectors $u_j$ of $A$ represent increasing frequencies as $j$ increases and the linear system satisfies the discrete Picard condition, i.e., the absolute value of the projections of the exact right-hand side $b^{\text{exact}}$ to the left singular subspaces of $A$ decays (on average) faster than the corresponding singular values. All given assumptions are natural for a broad class of ill-posed problems.

Based on the assumptions given above, it was shown in [4] how the noise contained in the right-hand side $b$ is propagated to the projected problem in the Golub-Kahan iterative bidiagonalization. Similar ideas are used in [7, 8, 3] for selection of a value of the regularization parameter for which the residual vector changes from being dominated by the remaining signal to being white-noise like. This leads to a parameter-choice method based on Fourier analysis of residual vectors. The noise propagation to reconstructed images computed by regularizing iterations is studied in [2].

In [4], information about the noise propagation is further used for estimating the unknown noise level in the data from the information available during the bidiagonalization process. The noise level detection is connected with the Gauss quadrature approximation of the Rieman-Stieltjes
distribution function determined by the input data. The presented estimate is then based on monitoring the absolute value of the first component of the left singular vector of the bidiagonal matrix corresponding to its smallest singular value. It can be computed at a negligible cost. Its accuracy and robustness is investigated using various test problems.

After reviewing the results presented in [4], we turn into the problem of approximating the unknown noise vector $b^{\text{noise}}$. Such information could further be used for improving accuracy of a regularized solution and construction of efficient stopping criteria for the Golub-Kahan iterative bidiagonalization. The work is in progress.

References


Communication-avoiding Krylov Subspace Methods

James Demmel, Michael Heroux, Mark Hoemmen, Maryggoob Mohiyuddin, and Katherine Yelick

Abstract

Krylov subspace methods (KSMs) for solving large, sparse linear systems and eigenvalue problems consume most of the run time of many scientific and engineering applications. Typical KSMs spend most of their time in computational kernels like sparse matrix-vector multiply (SpMV), dot products, and other vector-vector operations. These kernels all have performance dominated by communication, rather than floating-point arithmetic operations. “Communication” here means data movement, both between levels of a memory hierarchy (“sequential”) and between processors (“parallel”). Communication costs are orders of magnitude greater than arithmetic costs, and the gap is increasing rapidly for technological reasons. This suggests that for best performance, algorithms should minimize communication, even if that may require some redundant arithmetic operations. We call such algorithms “communication-avoiding.”

We have developed several communication-avoiding Krylov methods for solving linear systems and eigenvalue problems. Each of these methods computes the same approximate results in exact arithmetic as $s$ steps of its corresponding traditional KSM, but with the same communication costs as 1 step. We can do this given certain assumptions on the sparsity structure of the matrix, and for certain KSMs, both in theory (for many methods) and in practice (for some methods). Our new KSMs for solving linear systems include preconditioned and unpreconditioned versions of GMRES, CG, and BiCG. For solving eigenvalue problems (both $Ax = \lambda x$, for a nonsingular matrix $A$, and $Ax = \lambda Mx$, for nonsingular matrices $A$ and $M$), we have developed communication-avoiding versions of Arnoldi iteration, and both symmetric and nonsymmetric Lanczos iteration.

These algorithms, along with some implementations and performance results, are described in our doctoral dissertation [5]. Two collaborators have been extending our preliminary work in [5] on symmetric Lanczos iteration: Magnus Gustafsson, a doctoral student in scientific computing at Uppsala University in Sweden, and Ichitaro Yamazaki, a postdoctoral researcher at Lawrence Berkeley National Laboratory. Our University of California Berkeley collaborators Erin Carson and Nicholas Knight are currently working on improvements of our Communication-Avoiding BiCG algorithm.

Our communication-avoiding Krylov methods depend on two computational kernels: the “matrix powers kernel,” that can compute the results of $s$ calls to SpMV for the same communication cost as one call to SpMV, and Tall Skinny QR (TSQR), which can orthogonalize a collection of Krylov basis vectors even more accurately than competing orthogonalization methods, but with far fewer global reductions and memory traffic, and can also serve as an accurate rank-revealing decomposition. We present algorithms and implementations for the matrix powers kernel in [2, 3, 6]. Current work includes collaborations with the aforementioned Erin Carson and Nicholas Knight, and with Professor Michelle Mills Strout of Colorado State University and her graduate student Alan LaMielle, on developing and deploying optimized production-ready implementations of the matrix powers kernel. We discuss TSQR in great detail in [1], and have deployed an optimized hybrid-parallel implementation of TSQR in the Trilinos library of solvers [4]. We are currently collaborating with Michael L. Parks, senior technical staff at Sandia National Laboratories, on deploying TSQR as a faster orthogonalization kernel in iterative methods already implemented in Trilinos, such as the recycling GMRES solver GCRODR [7].

Coauthors and acknowledgments
My coauthor and doctoral dissertation advisor James Demmel is the Dr. Richard Carl Dehmel Distinguished Professor of the Electrical Engineering and Computer Science (EECS) Department of the University of California (UC) Berkeley, and is also Professor of Mathematics in the same. Katherine Yelick, my dissertation coadvisor, is Associate Director for Computing Sciences at Lawrence Berkeley National Laboratory (LBNL), the Director of the National Energy Research Scientific Computing Center (NERSC) at LBNL, and an EECS Professor at UC Berkeley. Marghoob Mohiyuddin, a doctoral student in the UC Berkeley EECS Department, developed new algorithms and implementations of the matrix powers kernel. Michael Heroux is a Distinguished Member of the Technical Staff at Sandia National Laboratories (SNL), and supervises my postdoctoral research.

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94AL85000. Special thanks to the United States Department of Energy Office of Advanced Scientific Computing Research (ASCR), for funding this research through the Extreme-Scale Algorithms and Architectures Software Initiative (EASI) project. During my graduate studies, my research was supported mainly by funding (Award #20080469) from Microsoft and Intel, with a matching grant from the University of California Discovery program (Award #DIG07-10227). I also received support from the ACM / IEEE High Performance Computing Fellowship Program,\textsuperscript{1} a student internship at Lawrence Livermore National Laboratory,\textsuperscript{2} and a National Science Foundation fellowship (2003 – 2005).

References


\textsuperscript{1}As of 2009, this fellowship has been named in memory of George Michael.

\textsuperscript{2}Lawrence Livermore National Laboratory is operated by Lawrence Livermore National Security, LLC, for the Department of Energy’s National Nuclear Security Administration.
Computations in Quantum Tensor Networks

Thomas Huckle, Thomas Schulte-Herbrüggen and Konrad Waldherr

Abstract

The computation of the ground state (i.e. the eigenvector related to the smallest eigenvalue) is an important task when dealing with quantum systems. As the dimension of the underlying vector space grows exponentially in the number of particles, one has to consider appropriate subsets promising both convenient approximation properties and efficient computations. The variational ansatz for this numerical approach leads to the minimization of the Rayleigh quotient. The Alternating Least Squares technique is then applied to break down the eigenvector computation to problems of appropriate size, which can be solved by classical methods. For efficient computations, we require fast computation both of the matrix-vector product and of the inner product of two decomposed vectors. To this end, appropriate representations of vectors are required.

In this talk we envisage several approaches which are already in use in physical quantum systems, such as Matrix Product States for one-dimensional systems and Projected Entangled Pair States for two-dimensional systems. We present these concepts from a mathematical point of view and show how computations (especially contraction schemes for the calculation of inner products) based on these concepts can be executed efficiently.

Furthermore, we consider classical tensor decomposition schemes and apply these concepts to our physical problem setting in order to develop a unifying method based on the ParaFac ansatz. It turns out that our ParaFac-like approach for the Rayleigh quotient minimization is meaningful and enables generalizations and modifications in several directions: it can be used in combination with mixed blocking structures, it can be applied in combination with Enhanced Line Search and it can easily be extended to higher dimensional problems.

Altogether, this talk will summarize different methods for the simulation of quantum many-body systems which enable time complexities that are polynomial in the number of particles.
A Subspace Shift Technique for Nonsymmetric Algebraic Riccati Equations

Bruno Iannazzo, Federico Poloni

Abstract

We consider the nonsymmetric algebraic Riccati equation (or NARE)

\[ 0 = R(X) := XCX - AX - XD + B, \]  

(1)

where \( X, B \in \mathbb{C}^{m \times n} \), \( A \in \mathbb{C}^{m \times m} \), \( C \in \mathbb{C}^{n \times m} \), \( D \in \mathbb{C}^{n \times n} \), and such that

\[ M = \begin{bmatrix} D & -C \\ -B & A \end{bmatrix} \]

is an M-matrix, either nonsingular or singular irreducible. In this case, Equation (1) is called M-NARE.

The solutions of the NARE (1) can be put in correspondence with certain \( n \)-dimensional invariant subspaces of the matrix

\[ H = \begin{bmatrix} D & -C \\ B & -A \end{bmatrix}. \]  

(2)

More precisely, a matrix \( X \in \mathbb{C}^{m \times n} \) is a solution of (1) if and only if

\[ H \begin{bmatrix} I_n \\ X \end{bmatrix} = \begin{bmatrix} I_n \\ X \end{bmatrix} \begin{bmatrix} D - CX \end{bmatrix}. \]  

(3)

Thus, the spectral properties of the matrix \( H \) are relevant for the analysis of the NARE (1).

For an M-NARE, where \( H = JM \) for \( J = \text{diag}(I_n, -I_m) \), it can be proved (see [3] and the references therein) that the eigenvalues of \( H \) can be ordered by non increasing real part such that

\[ \text{re}(\lambda_1) \geq \cdots \geq \text{re}(\lambda_{n-1}) > \lambda_n \geq 0 \geq \lambda_{n+1} > \cdots \geq \text{re}(\lambda_{m+n}), \]

that is, \( n \) eigenvalues belong to the closed right half complex plane and the other eigenvalues to the closed left half plane, and the central eigenvalues, \( \lambda_n \) and \( \lambda_{n+1} \), are real and separated from the other eigenvalues.

Moreover, the matrix \( H \) associated with an M-NARE has a unique \( n \)-dimensional invariant subspace corresponding to the \( n \) rightmost eigenvalues, namely \( \lambda_1, \ldots, \lambda_n \), which we call the \( n \)-dimensional semi-unstable invariant subspace of \( H \). In the applications [4, 5], the required solution of the M-NARE is the one for which the columns of \( \begin{bmatrix} I_n \\ X \end{bmatrix} \) span the semi-unstable invariant subspace of \( H \).

The existence of this solution has been proved and it turns out to be the minimal element-wise nonnegative solution of (1) (see [2]).

Equation (1) is usually solved either by some matrix iteration, e.g., the Cyclic Reduction (CR) or the Structure-preserving Doubling Algorithm (SDA), whose limits yield the required solution or using the ordered Schur form of \( H \).

Both the conditioning of the equation and the convergence speed of the iterations are strictly related to the relative gap between the central eigenvalues of \( H \), i.e., \( \text{gap} = |\lambda_n - \lambda_{n+1}| / \|H\| \).

If \( \lambda_n = \lambda_{n+1} (= 0) \), then the minimal nonnegative solution of equation (1) is ill-conditioned and the convergence of iterations such as CR and SDA, which is quadratic in the generic case, turns
to linear. We speak of critical case, since in these cases the required solution $X$ is critical, namely $\mathcal{R}'(X)$ is singular, where $\mathcal{R}'(X)$ is the Fréchet derivative of the operator $\mathcal{R}(X)$.

In such cases, the shift technique of [3] has proved to be useful. It consists in making a special rank-one correction of $H$, obtaining a new Riccati equation with the same minimal solution. The new equation has better conditioning and the convergence of iterations is quadratic again.

However, ill-conditioning and slow convergence appear also in the close-to-critical case, that is when the derivative of $\mathcal{R}(X)$ is near to a critical matrix, or, in terms of the central eigenvalues, when $\lambda_n$ and $\lambda_{n+1}$ both approach 0. This is the worst situation since the numerical solution of the matrix equation is problematic and the use of the shift technique is not recommended since it relies on the computation of an ill-conditioned eigenvector.

We present a new technique to handle the close-to-critical case. The technique relies on the fact that, for the M-NARE, there exists a unique 2-dimensional invariant subspace of $H$ associated with the central eigenvalues $\lambda_n$ and $\lambda_{n+1}$, which we call the 2-dimensional central invariant subspace of $H$.

If the central eigenvalues $\lambda_n$ and $\lambda_{n+1}$ of $H$ are close to each other but well separated from the other eigenvalues of $H$, then, while the eigenvectors corresponding to $\lambda_n$ and $\lambda_{n+1}$ are ill-conditioned, the 2-dimensional central invariant subspace shows good conditioning.

The proposed technique is based on a rank-two modification of $H$ made by means of the central invariant subspace. We move the central eigenvalues together and we obtain a new NARE, with the same solution as the original or a solution which is a rank-one modification of the solution of the original equation. The new equation has better conditioning and, in certain cases, the convergence of iterations like the SDA is much faster.

Finally, the extension to reducible M-matrices and the applicability of the SDA in the shifted and double shifted case are discussed.

References


Abstract

We consider the problem of randomized matrix multiplication, and in particular a Monte Carlo algorithm by Drineas, Kannan and Mahoney. We present improved probabilistic bounds on the accuracy of this algorithm due to randomization. We also present probabilistic bounds for the sensitivity of the algorithm to changes in the inputs. Numerical experiments indicate that the bounds are realistic and hold even for moderate matrix dimensions.
Solving Nonlinear Eigenvalue Problems with a Linear Arnoldi Method in a Function Setting and a Characterization of Associated Invariant Pairs

Elias Jarlebring, Wim Michiels, Karl Meerbergen

Abstract

Let $B(\lambda) \in C^{n \times n}$ be a parameter-dependent matrix with elements depending analytically on $\lambda \in C$ and consider the nonlinear eigenvalue problem (NEP),

$$\lambda B(\lambda)x = x,$$

where $x \in C^n \{0\}$. This is an algebraic reformulation of the standard NEP, $M(\lambda)x = 0$, and it will simplify our notation.

We now associate the NEP (1) with an operator $B$ acting on vectors of functions of the variable $\theta \in R$. The action is defined as integration of the vector of functions and the addition of a constant vector,

$$(B\varphi)(\theta) := \int_0^\theta \varphi(\theta) \, d\theta + \left( B(\frac{d}{d\theta})\varphi \right)(0).$$

See [1, Definition 1] for a precise definition. This operator is constructed such that it is equivalent to the nonlinear eigenvalue problem (1) in the sense that the reciprocal eigenvalues of the operator $B$ are the solutions (1). More precisely, the following statements are equivalent [1, Theorem 1].

i) The pair $(\varphi, \lambda)$ is a solution to $\lambda B\varphi = \varphi$ with $\varphi(\theta) = xe^{\lambda\theta}$.

ii) The pair $(x, \lambda)$ is a solution to (1).

From the above equivalence, we expect that the Arnoldi method in a function setting for the operator $B$ converges to solutions of (1) and favors solutions close to the origin. In [1] we apply the Arnoldi method in a function setting with polynomials, e.g., the Chebyshev polynomials, and carry out iterations with functions of the form,

$$\varphi(\theta) = (x_0, \ldots, x_k) \begin{pmatrix} \hat{T}_0(\theta) \\ \vdots \\ \hat{T}_k(\theta) \end{pmatrix}, \quad x_0, \ldots, x_k \in C^n,$$

where $\hat{T}_i, i \in N$ are the Chebyshev polynomials. We show how we can carry out the algorithm in an exact way (without approximation) with standard linear algebra operations involving matrices (not operators) acting on the coefficients $(x_0, \ldots, x_k)$. Due to the fact that the algorithm can be interpreted as the Arnoldi method in a function setting, the convergence and other properties are expected to be the same as for the standard Arnoldi method. In [1] we use the method to solve large-scale nonlinear eigenvalue problems in a reliable fashion.

The Arnoldi method for standard eigenvalue problems has been considerably improved by using theory for restarting of the iteration. Several restarting strategies for standard eigenvalue problems can be interpreted as a nesting of the Arnoldi method with an outer iteration which converges to an invariant pair of the eigenvalue problem. We now make first step towards constructing an adaption of the Arnoldi algorithm in [1] allowing efficient restarting. We show how to represent an invariant
pair of the operator $\mathcal{B}$ by making a connection with a result on invariant pairs for nonlinear linear eigenvalue problems recently presented in a somewhat different setting [2].

Consider a (for the moment) arbitrary matrix $S \in \mathbb{C}^{n \times n}$ and a vector $c \in \mathbb{C}^n$. Let $\psi_c$ and $\varphi_c$ be two functions with the structure,

$$
\psi_c(\theta) := (y_1, \ldots, y_p) \exp(\theta S) c \quad \text{and} \quad \varphi_c(\theta) := (y_1, \ldots, y_p) S \exp(\theta S) c.
$$

with $y_1, \ldots, y_p \in \mathbb{C}^n$. The following statements are equivalent.

i) For all $c \in \mathbb{C}^n$, the functions $\psi_c$ and $\varphi_c$ fulfill,

$$
\mathcal{B} \varphi_c = \psi_c.
$$

ii) The pair $(S, (y_1, \ldots, y_p))$ is an invariant pair of the nonlinear eigenvalue problem (1) in the sense of [2, Definition 1].

Consequently, from [2, Lemma 1] we have that eigenvalues of $S$ are solutions to the NEP (1) if $S$ fulfills (3) and (4).

In order to represent an approximation of the invariant pair, we now adapt the algorithm in [1] by extending the considered structure of the functions by (matrix) exponential functions. We work with functions which are more general than (2) and of the form,

$$
\varphi(\theta) = (\tilde{x}_1, \ldots, \tilde{x}_p) \exp(\theta \tilde{S}) c + (x_0, \ldots, x_k) \begin{pmatrix} \hat{T}_0(\theta) \\ \vdots \\ \hat{T}_k(\theta) \end{pmatrix}, \quad \tilde{x}_1, \ldots, \tilde{x}_p \in \mathbb{C}^n,
$$

where $\tilde{S}$ is a matrix typically approximating the matrix $S$ solving (4). A key to carrying out the iteration is that the action of a function with the structure (5) can, similar to [1], also be carried out by standard linear algebra manipulations of coefficients. This stems from the fact that if we apply $\mathcal{B}$ to an approximation of the invariant pair, we still have an exponential structure corrected by a constant vector. That is,

$$
(\mathcal{B} \varphi)(\theta) = (\tilde{x}_1, \ldots, \tilde{x}_p) \tilde{S}^{-1} \exp(\theta \tilde{S}) c + z \quad \text{with} \quad \varphi(\theta) = (\tilde{x}_1, \ldots, \tilde{x}_p) \exp(\theta \tilde{S}) c,
$$

and $z$ is a constant vector computable (with the help of matrix functions) from $\tilde{x}_1, \ldots, \tilde{x}_p$ and $\tilde{S}$.

The restarting procedure in [3] involves a partial Schur decomposition of the Hessenberg matrix. We illustrate how we can after $k$ iterations restart the iteration with $p$ functions with the structure (5) and study the impact of the choice of $c$, which can be interpreted as the representation of the linear combination of the approximations of the Schur functions to be used in the restart.

References


A Posteriori Error Estimates including Algebraic Error and Stopping Criteria for Iterative Solvers

Pavel Jiránek, Zdeněk Strakoš, and Martin Vohralík

Abstract

In numerical solution of partial differential equations, a natural question is whether the computed result is a sufficiently accurate approximation of the exact solution of the problem at hand. A posteriori error estimates aim at giving an answer to this question by providing computable upper bounds on a norm of the difference between the approximate finite-dimensional and the exact solution. From a large amount of literature on this subject we point out, e.g., the books [1, 3].

Discretization of a partial differential equation using, e.g., finite element, finite difference, and finite volume methods leads to a system of linear algebraic equations which is usually sparse and large with the dimension depending on the resolution of the underlying grid. Moderately sized systems of linear equations can be solved by using direct methods, typically variants of Gaussian elimination. For large systems, preconditioned iterative methods become competitive and with increasing size they represent the only viable alternative. Iterative methods can produce, in contrast with direct ones, an approximation to the solution at each iteration step with the amount of computational work depending on the number of performed iterations. An efficient PDE solver should use this principal advantage by stopping the algebraic solver whenever the algebraic error drops to the level at which it does not affect the overall error. This requires evaluation and comparison of algebraic and total errors, which represents a conceptually open problem.

The difficulty can be illustrated on an example of the Galerkin finite element discretization of elliptic problems. Here the energy norm of the overall error can be related to the energy norms of the discretization and algebraic errors by the equality \[ \|p - p_h\|_2^2 = \|p - p_h^a\|_2^2 + \|p_h - p_h^a\|_2^2 \] (see, e.g., [4, Section 2.1]), where \(p\) is the exact (weak) solution of the partial differential equation, \(p_h\) is the exact solution of the discretized problem, and \(p_h^a\) its approximation. Even though this equality elegantly separates the discretization and algebraic parts of the overall error, it can barely be used in practical stopping criteria since the exact solution of the discrete problem is not known (the algebraic error can be estimated using purely algebraic tool).

The simplest and frequently used stopping criterion is based on the evaluation of the relative Euclidean norm of the residual vector. It is however only roughly connected with the size of the overall error in the continuous problem and usually not even this connection is considered. Consequently, one either continues the algebraic iterations until the residual norm is not further reduced possibly without getting any further improvement of the overall error or one stops earlier with a risk that the computed approximation is not sufficiently accurate.

In [6] we consider a second-order elliptic (pure diffusion) model problem \(-\nabla \cdot (S \nabla p) = f\) in a domain \(\Omega\), where \(p\) is the unknown function defined on \(\Omega\), \(S\) is a (diffusion) tensor, and \(f\) is a source term. On the boundary of \(\Omega\) we prescribe an inhomogeneous Dirichlet condition. For the discretization of the problem, we consider a conforming triangulation (in 2D) of \(\Omega\) and a general locally conservative cell-centered finite volume scheme; see, e.g., [5]. By solving the resulting discrete problem, we obtain a discontinuous element-wise quadratic approximation \(\tilde{p}_h\) to the function \(p\) with continuous inter-element diffusion fluxes. The discretized problem is, however, not solved exactly (on purpose). Instead of \(\tilde{p}_h\) we get an approximation \(\tilde{p}_h^a\) using, e.g., the preconditioned conjugate gradient method.
Apart from few exceptions, existing a posteriori error estimates rely on the assumption that the linear algebraic system resulting from the discretization is solved exactly. Our first goal is to derive an a posteriori error estimate which takes into account an inexact solution of the linear algebraic system associated with the discretized partial differential equation. For this purpose we extend the results of [7]. The obtained upper bound for the energy norm of the error is the sum of three terms:

$$\|p - \tilde{p}_h\| \leq \eta_{NC} + \eta_O + \eta_{AE}.$$

Here the estimator $\eta_{NC}$ measures the nonconformity of the approximate solution, which reflects mainly the discretization error, the estimator $\eta_O$ corresponds to the interpolation error in the approximation of the source term, which turns out to be of higher-order, and the algebraic error estimator $\eta_{AE}$ bounds the error of the inexact solution of the linear algebraic system.

Comparing the nonconformity and algebraic error estimates, the second goal is to construct, in the context of our model problem, efficient stopping criteria for iterative algebraic solvers. The approach proposed previously in [2] uses a priori error bounds on the discretization error. Here we use a posteriori error estimates and require that the algebraic error estimate $\eta_{AE}$ is a sufficiently small multiple of the nonconformity estimate $\eta_{NC}$. We can also establish the efficiency of the estimate. Indeed, it represents a lower bound for the true error up to a generic constant independent on the discretization parameter. In addition, an element-wise comparison of the discretization and algebraic error estimates ensures the (stronger) local efficiency. We can thus use the estimates for an efficient prediction of the overall error size and its distribution amongst the mesh elements. Therefore the proposed estimates are suitable for adaptive mesh refinement which takes into account the inaccuracy of the algebraic computations.

Extension of the described results to more general problems and other discretization methods is under investigation.

References


Stratification of Full Normal Rank Polynomial Matrices

Stefan Johansson, Bo Kågström, and Paul Van Dooren

Abstract

To study how a polynomial matrix

\[ P(s) := P_d s^d + \ldots + P_1 s + P_0, \quad \text{where} \quad P_i \in \mathbb{C}^{m \times n}, \]

beha
s under small perturbations is a critical task, since computing the structural elements (elementary divisors and minimal indices) of \( P(s) \) is sensitive to small perturbations in the coefficient matrices \( P_i \) [7]. It is therefore important to know how small variations of the coefficients \( P_i \) can change the computed canonical structure. Our approach to such an analysis is to study perturbations of linearizations of polynomial matrices.

Polynomial matrices arise, for example, from dynamical systems described by sets of differential equations with constant coefficient matrices

\[ P_d x^{(d)}(t) + \ldots + P_1 x^{(1)}(t) + P_0 x(t) = f(t), \]

where \( x(\cdot) \in \mathbb{C}^n, f(\cdot) \in \mathbb{C}^m \), and \( x^{(i)}(t) \) is the \( i \)-th derivative of the solution vector \( x(t) \). Taking the Laplace transform of this equation and imposing zero initial conditions, yields the differential-algebraic equation

\[ P(s) \hat{x}(s) = \hat{f}(s), \]

where \( \hat{x}(\cdot) \) and \( \hat{f}(\cdot) \) are the Laplace transforms of \( x(t) \) and \( f(t) \), respectively. Currently, we restrict our study to polynomial matrices with full normal rank \( r \) (revealed by the Smith normal form) is equal to the rank of \( P(s) \) at any value of \( s \in \mathbb{C} \) which is not a zero of \( P(s) \). We also consider \( P(s) \hat{x}(s) = \hat{f}(s) \), when \( P(s) \) is monic, i.e., \( P(s) \) is \( n \times n \) with \( P_d \equiv I_n \).

One tool that can be used to analyze the qualitative information about nearby systems is the theory of stratification [2, 1, 6]. A stratification reveals the closure hierarchy of orbits and bundles of nearby canonical structures and give important qualitative information about the underlying dynamical system. It shows which canonical structures can be reached by a small perturbation and the relation among these structures. A stratification is represented as a graph where each node represents an orbit or bundle of a canonical structure and an edge a covering relation. The cover relations reveal the nearest neighbors in the closure hierarchy and are expressed as combinatorial rules acting on integer sequences representing a subset of the canonical structural elements.

We extend the theory of stratification for matrix pencils [1, 2] and matrix pairs [3, 6] to polynomial matrices by making use of appropriate linearizations. It is well known that the right linearization (called a companion form) of a polynomial matrix \( P(s) \) is

\[ sB_r + A_r := s \begin{bmatrix} I_m \\ \vdots \\ I_m \\ P_d \end{bmatrix} + \begin{bmatrix} 0 & & & P_0 \\ -I_m & \ddots & \ddots & \vdots \\ \vdots & \ddots & 0 & \vdots \\ -I_m & \ddots & \ddots & P_{d-1} \end{bmatrix}. \]

The matrix pencil \( sB_r + A_r \) is a first order (or linear) polynomial matrix, which has the same structural elements as the polynomial matrix \( P(s) \) with normal rank \( r = m \). Moreover, the special
structure of this pencil must be taken into account when analyzing perturbations of such linearizations.

We show that the stratification rules for linearizations of full normal rank polynomial matrices $P(s)$ are a subset of the rules for general rectangular matrix pencils. In addition, we show that the stratification rules for linearizations of $P(s)\dot{x}(s) = \dot{f}(s)$ can be expressed as a subset of the rules for controllability pairs $(A, B)$ associated with linear time-invariant systems. The theory of stratification is illustrated and the qualitative information is analyzed on examples of dynamical systems. The stratification graphs are generated with StratiGraph [4, 5], which is a software tool for determining and presenting orbit and bundle closure hierarchies. This is work in progress.

References


Efficient and Reliable Algorithms for Challenging Matrix Computations targeting Multicore Architectures and Massive Parallelism

Bo Kågström

Abstract

Along with the evolution towards massively parallel HPC systems with multicore nodes, there is an immense demand of new and improved scalable, efficient, and reliable numerical algorithms and library software for fundamental and challenging matrix computations. Such algorithms and software are used as building blocks for solving current and future very large-scale computational problems in science and engineering.

Recently, the Umeå research group has presented several novel results concerning challenging matrix computations, including (1) Parallel and cache-efficient in-place matrix storage format conversion [5]; (2) Parallel two-stage reduction to Hessenberg form using shared memory [7, 8]; (3) Parallel QR and QZ multishift algorithms with advanced deflation strategies [3, 6]; (4) Parallel eigenvalue reordering in real Schur forms [4]; (5) The SCASY library - parallel solvers for Sylvester-type matrix equations with applications in condition estimation [1, 2].

Topic (1) concerns techniques and algorithms for efficient in-place conversion between standard and blocked matrix storage formats. Such functionality enables numerical libraries to use various data layouts internally for matching blocked algorithms and data structures to a memory hierarchy.

Topics (2)–(5) concern the solution of large-scale dense eigenvalue problems and matrix equations via two-sided transformation methods and provide new functionality for scalable HPC computations. Fundamental two-sided matrix computations include the parallel reduction to Hessenberg and Schur forms as well as eigenvalue reordering. Altogether this makes it possible to compute invariant and deflating subspaces associated with a specified spectrum, e.g., all eigenvalues inside the unit circle, along with error estimates of computed quantities. Compared to one-sided factorizations (e.g., LU, Cholesky, QR), these two-sided matrix decompositions have much more complex data dependencies that pose challenging parallel computing problems. A key technique to cope with these problems is our parallel multi-window technique combined with delaying of matrix updates and the accumulation of transformations. As an example, our novel parallel QR algorithm is one to two orders of magnitude faster than the existing ScaLAPACK routine.

In this presentation, we will review and highlight some of our most recent results concerning the topics (1)–(5). These results represent collaborative efforts together with several people including Björn Adlerborn, Robert Granat, Fred Gustavson, Lars Karlsson, Daniel Kressner, and Meiyue Shao.

References


Refining the General Symmetric Eigenproblem

W. Kahan

Abstract

Refining the General Symmetric Eigenproblem

Real symmetric matrices $A$ and $H$, the latter also positive definite, define eigenproblem $Av = \lambda Hv$ whose every eigenvalue $\lambda$ is the square of a vibrational frequency of an undamped elastic structure. For over forty years engineers have computed values $\lambda$ and vectors $v$, when $A$ and $H$ are sparse but their Cholesky factors aren’t, by an iterative process similar to the Jacobi iteration used when $H = I$; see ch. 15 of B.N. Parlett’s book “The Symmetric Eigenproblem” (1998), and a paper by Slapnicar and Hari (1991). The iteration is known to converge quadratically if it starts close enough to its goal but global convergence was an open question unless $H = I$. Herein global convergence is proved from every starting situation; the proof is unusual in so far as it exploits the hitherto unnoticed monotonicity of a determinant during the iteration.

Numerical instability threatens this eigenproblem in two ways. One of them can cause an iteration-step to spoil results unless it uses an unobvious formula in its implementation. The second way arises when $H$ is too nearly singular, especially if $H$ shares a near-nullspace with $A$. In this case eigenvalues otherwise well-determined by the data can be corrupted by roundoff unnecessarily, according to a novel error-analysis, unless computed eigenvectors are refined in a not-too-costly auxiliary computation. For details see <www.eecs.berkeley.edu/~wkahan/Math128/GnSymEig.pdf>

The iteration and error-analyses are under test. Among test data are segments of Hilbert matrices for which Matlab programs are supplied to compute quickly four Cholesky factors (upper $U'U$ and lower $L'L$) of these Hilbert matrices and their inverses, all with errors amounting to at most about three units in each element’s last sig. bit because integer-valued floating-point arithmetic is used extensively. Likewise for bidiagonal matrices whose squared singular values are the eigenvalues $\lambda$ for data $A$ and $H$ consisting of adjacent segments of Hilbert matrices. For details see <www.eecs.berkeley.edu/~wkahan/MathH110/HilbMats.pdf>.

---

Parallel Two-Stage Reduction to Hessenberg Form using Shared Memory

Lars Karlsson and Bo Kågström

Abstract

The Hessenberg form of a square matrix \( A \in \mathbb{R}^{n \times n} \) is an upper Hessenberg matrix \( H \in \mathbb{R}^{n \times n} \) that is orthogonally similar to \( A \), i.e., there is an orthogonal matrix \( Q \in \mathbb{R}^{n \times n} \) such that \( H = Q^T AQ \). The Hessenberg form is commonly used as a preprocessing step in solvers for the nonsymmetric eigenvalue problem, e.g., the QR algorithm, and in solvers for Sylvester-type matrix equations, e.g., the Bartels-Stewart method and the Hessenberg-Schur method.

A Hessenberg reduction algorithm computes the Hessenberg form \( H \) of \( A \) via a sequence of two-sided transformations applied to \( A \). The transformations are applied alternatingly to the left-hand and right-hand sides of \( A \), which makes it difficult to effectively use the memory hierarchy and parallel processing. For example, the prevailing Hessenberg reduction algorithm reduces the columns of \( A \) from left to right using Householder reflections. Performance is enabled by accumulation of reflections into compact WY representations which are applied using matrix-matrix multiplications.

The standard blocked algorithm above is far superior to the corresponding unblocked algorithm, but it still performs roughly 20% of the floating point operations inside large matrix-vector multiplications. This limits the performance in practice since large matrix-vector multiplications are inherently communication-intensive and non-scalable on current multicore processors.

We have recently studied a two-stage algorithm for Hessenberg reduction that reduces the matrix to an intermediate form that has \( r \) nonzero subdiagonals and then further to actual Hessenberg form \([1, 2]\). Two-stage algorithms are common for reduction of symmetric matrices since the second stage is cheap compared to the first stage. However, the second stage cannot be neglected for the (nonsymmetric) Hessenberg reduction.

A straightforward generalization of the standard blocked Hessenberg reduction algorithm leads to an algorithm for the first stage that performs roughly all of its floating point operations inside matrix-matrix multiplications. We present our recent parallel implementation that is based on coarse-grained dynamic scheduling \([2]\).

The second stage of the reduction is accomplished by a well-known bulge-chasing procedure which is based on Householder reflections. The columns are reduced from left to right and each column reduction introduces a sequence of small bulges. The first column of each such bulge must be reduced before the next column reduction. We recently showed how to add blocking for the memory hierarchy and coarse-grained parallelism to this algorithm \([1]\). The blocked algorithm generates all Householder reflections from multiple consecutive sweeps while updating only the entries within a band around the main diagonal. The delayed updates are applied in parallel using a cache-efficient scheme. One level of look-ahead further improves the parallel performance by eliminating overhead caused by the sequential part of the algorithm.

Computational experiments show that our algorithm for the first stage is significantly faster than a corresponding algorithm which uses multi-threaded BLAS instead of coarse-grained dynamic scheduling. Furthermore, the blocking and parallelization of the second stage is shown to improve the performance of the bulge-chasing algorithm by an order of magnitude \([1]\). Finally, the full two-stage Hessenberg reduction algorithm attains a speedup of roughly 2.5 with respect to the current LAPACK implementation of the standard blocked (one-stage) algorithm \([2]\).
References


Modifications of an Algorithm for Factoring Symmetric Banded Indefinite Matrices

Linda Kaufman

Abstract

In 2007, the author outlined an algorithm for factoring a banded symmetric, indefinite matrix based on version D of the Bunch Kaufman Algorithm using a sequence of 1 x 1 and 2 x 2 pivots. The algorithm preserves symmetry and band structure and limits element growth in the factorization. For an \( n \times n \) symmetric matrix of bandwidth \( 2m + 1 \), the algorithm requires between \( \frac{1}{2}nm^2 \) and \( \frac{5}{4}nm^2 \) multiplications and at most \( (2m + 1)n \) locations compared to nonsymmetric Gaussian elimination which requires between \( nm^2 \) and \( 2nm^2 \) multiplications and at most \( (3m + 1)n \) locations.

For a sequence of 1x1 pivots, there is no fill-in outside the band and one can easily create a block version using level 3 blas. However for a 2 x 2 pivot, there is a rank 1 fill-in wing outside the band which naturally one might eliminate using a sequence of stabilized elementary transformations or rotations (i.e. level 1 Blas). We show how one can overlap sequences to accelerate the algorithm. We also describe a burn at both ends or twisted algorithm to decrease the element growth and the size of the wings.
Approximations of Third Order Tensors as Sums of (Non-negative) Low-rank Product-Cyclic Tensors

Misha E. Kilmer, and Tamara G. Kolda

Abstract
Operations with tensors, or multiway arrays, have become increasingly prevalent in recent years. The list of applications involving operations on tensors is lengthy and includes psychometrics, signal processing and data mining, to name but a few. A common feature in such applications is the need to derive a compressed representation of the data that takes advantage of the multidimensional structure, since collapsing multiway data to matrices often has undesirable consequences.

Traditionally, tensors are decomposed using CANDECOMP/PARAFAC (CP), Tucker, or some variation thereof. In this talk, we explore a novel decomposition that shows promise with respect to the tensor approximation problem. Specifically, we decompose a third order tensor as a finite sum of specially structured tensors, which we call product-cyclic tensors (PCTs). Each PCT is defined completely in terms of a pair of matrices, \(G\) and \(H\), with \(G \in \mathbb{R}^{\ell \times n}\), \(H \in \mathbb{R}^{m \times n}\) such that the \(k\)th frontal slice of the PCT is given by \(GZ^{k-1}H^T\), where \(Z\) denotes the \(n \times n\) downshift matrix. Notationally, we represent such a tensor as

\[ G \diamond H \in \mathbb{R}^{\ell \times m \times n}, \quad (G \diamond H)_{:,i,k} = GZ^{k-1}H^T. \]

It can be shown (see below) that any \(\ell \times n \times m\) real tensor \(A\) can be written as a sum of at most \(\min(\ell, n, m)\) PCTs. In this talk, we focus on optimally approximating a tensor as a sum of fewer than \(\min(\ell, n, m)\) PCTs while adding constraints, such as non-negativity constraints or maximal matrix rank constraints, on the pairs of matrices that comprise each of the PCTs. This approach leads us to a new way of generating compressed representations of third order tensors.

The motivation for a PCT-based decomposition arises from recent work [1, 2] which introduces a framework for tensors built around a \(t\)-product. The \(t\)-product \(A \ast B\) between an \(\ell \times k \times n\) tensor \(A\) and \(k \times m \times n\) tensor \(B\) is an \(\ell \times m \times n\) tensor. That tensor results from first computing the matrix product \(\text{circ}(A) \cdot \text{unfold}(B)\), where \(\text{unfold}(B)\) is the \(kn \times m\) block matrix formed by unstacking \(B\) by frontal faces and \(\text{circ}(A)\) is the \(\ell n \times kn\) block circulant matrix with \(\text{unfold}(A)\) as the first block column, and then folding the resultant matrix product back into an \(\ell \times m \times n\) tensor. If \(A \in \mathbb{R}^{\ell \times m \times n}\), it has been shown that there exists an orthogonal \(U \in \mathbb{R}^{\ell \times \ell \times n}\) (i.e. \(U^T U = I\) where \(I_{:,1} = I, I_{:,j} = 0, j \neq 1\)), an orthogonal \(V \in \mathbb{R}^{m \times m \times n}\) and \(S \in \mathbb{R}^{\ell \times m \times n}\) with diagonal frontal faces such that

\[ A = U \ast S \ast V^T = \sum_{i=1}^{\min(\ell, n)} U_{:,i,\ast} \ast S_{i,\ast,\ast} \ast V_{:,i,\ast}^T, \quad (1) \]

where the transpose of a third order tensor \(V\) is the third order tensor that results from transposing each frontal face of \(V\), then reordering those faces 2 through \(n\) in reverse order. This decomposition is referred to as the t-SVD [2].

A visual interpretation of the sum in (1) is worthwhile. The expression \(S_{i,\ast,\ast}\) refers to a tube of \(S\), and it is appropriate to think of such tubes as playing the role of scalars (see [3]). For each fixed \(i\), note that \(U_{:,i,\ast}\) is a length \(\ell\) column vector of tubes of length \(n\) (i.e. a matrix twisted into the page) and \((S_{i,\ast,\ast} \cdot V_{:,i,\ast}^T)\) is a length \(m\) row vector whose entries are tubes of length \(n\) (also a matrix when orientated appropriately).
Thus, pairs of matrices play an important role in the factorization. Let $G$ be a matrix of size $\ell \times n$, and use $\text{twist}(G)$ to denote the equivalent $\ell \times 1 \times n$ tensor such that entry $(i, 1, j)$ is $g_{ij}$. Now

$$G \diamond H \equiv \text{twist}(G) \ast \text{twist}(H)^T.$$ 

It is shown in [2], that a partial t-SVD gives an optimal approximation in the Frobenius norm in the sense that for $k \leq \min(\ell, m)$:

$$\|A - \sum_{i=1}^{k} U_{i,:} \ast S_{i,:} \ast V_{i,:}^T\|_F = \min_{G_i \in \mathbb{R}^{\ell \times n}, H_i \in \mathbb{R}^{m \times n}} \|A - \sum_{i=1}^{k} G_i \diamond H_i\|_F.$$ 

In that paper, a tensor compression algorithm is proposed based on a combination of a partial t-SVD of the tensor in combination with partial matrix SVDs of appropriately oriented slices $U_{i,:}$ and $S_{i,:} \ast V_{i,:}^T$. The result is an approximate CP decomposition of the tensor, which they show is useful in an application in image deblurring.

However, finding the t-SVD directly as proposed in [2] is expensive because the factorization process requires taking Fourier transforms along all tubes of $A$. Furthermore, using the t-SVD as a first step in deriving a CP approximation to $A$ does not allow the enforcement of constraints, such as non-negativity, directly on the terms in the sum.

Therefore, we consider instead the problem of finding an approximation to $A$ as a sum of $k$ matrix outer products, subject to constraints on the matrices:

$$\min_{G_i \in \mathbb{R}^{\ell \times n}, H_i \in \mathbb{R}^{m \times n}} \|A - \sum_{i=1}^{k} G_i \diamond H_i\|_F, \quad \text{subject to constraints on } G_i, H_i, i = 1, \ldots, k.$$ 

Specifically, we consider as constraints

- restricting the matrix rank of $G_i, H_i$. In this case, the first order necessary conditions lead to sequences of matrix eigenvalue problems. We show that the output can be used to construct a low rank (in the tensor sense) CP approximation to $A$.
- adding non-negativity constraints to the matrices in the decomposition, which seems to lead to sparsity in the matrices in the decomposition.

The algorithms are of the alternating least squares (ALS) variety, but what makes them potentially superior to traditional ALS-type algorithms for computing CP decompositions is the fact that the alternation is only over a pair of matrices as opposed to the usual three for a third order tensor, making the convergence easier to analyze and the convergence behavior, it appears, quite robust. We give numerical results which show the advantages of our approach.

References


Incomplete Cyclic Reduction for Narrow Banded and Diagonally Dominant Linear Systems

Carl Christian Kjelgaard Mikkelsen, and Bo Kågström

Abstract

The current implementation of ScaLAPACK contains two routines for solving linear systems which are narrow banded and strictly diagonally dominant by rows [7, 1]. The algorithm is complete block cyclic reduction corresponding to a particular block partitioning of the matrix. Let \( A \in \mathbb{R}^{n \times n} \) be a block tridiagonal matrix,

\[
A = \begin{bmatrix}
  D_1 & F_1 & & \\
  E_2 & \ddots & \ddots & \\
  & \ddots & \ddots & F_{m-1} \\
  & & E_m & D_m
\end{bmatrix},
\]

which is strictly diagonally dominant by rows. For the sake of simplicity, we assume that \( m = 2^k - 1 \) for an integer \( k > 0 \). Let \( D \) denote the main block diagonal of \( A \), i.e.

\[
D = \text{diag}(D_1, D_2, \ldots, D_m)
\]

and consider the auxiliary matrix \( B \) given by

\[
B = D^{-1}(A - D).
\]

Then, it follows readily, that if

\[
f \neq 0, \quad Ax = f, \quad Dy = f,
\]

then

\[
\frac{\|x - y\|_\infty}{\|x\|_\infty} \leq \|B\|_\infty.
\]

In short, the significance of the off diagonal blocks can be measured using the auxiliary matrix \( B \). D. Heller [2] showed that if \( A \) is strictly diagonally dominant by rows, then block cyclic reduction is well defined and yields a sequence of block tridiagonal matrices,

\[
A = A^{(0)}, A^{(1)}, A^{(2)}, \ldots, A^{(j)}, \quad j < k,
\]

such that \( A^{(j)} \) has \( m_j = 2^{k-j} - 1 \) diagonal blocks and the corresponding auxiliary matrices \( B^{(j)} \) given by

\[
B^{(j)} = D^{(j)-1}(A^{(j)} - D^{(j)})
\]

satisfy

\[
\|B^{(j+1)}\|_\infty \leq \|B^{(j)}\|_\infty^2.
\]

If \( A \) is strictly diagonally dominant by rows, then the dominance factor is defined by

\[
\epsilon = \max_i \left\{ \frac{1}{|a_{ii}|} \sum_{j \neq i} |a_{ij}| \right\} \in [0, 1).
\]
We have shown that it is straightforward to incorporate the dominance factor $\epsilon$ into the analysis. Specifically, we have

$$\|B^{(j)}\|_\infty \leq \epsilon^{2^j}$$

and this estimate is tight [6]. Naturally, this estimate carries to the banded case, but it does not reflect the banded structure.

Let $A$ be a narrow banded matrix which is also diagonally dominant by rows and let $k$ denote the half bandwidth of $A$, i.e.

$$|i - j| > k \Rightarrow a_{ij} = 0.$$  

In the ScaLAPACK implementation, the odd numbered diagonal blocks have size $\mu = qk$ for some large integer $q$, while the even numbered diagonal blocks have dimension $k$. We have shown [6], that dropping the off diagonal blocks in the initial Schur complement, is equivalent to solving a perturbed linear system

$$(A + \Delta A)x = f$$

for which

$$\|\Delta A\|_\infty \leq \epsilon^{1+q} \|A\|_\infty.$$  

This estimate is closely related to our recent analysis of the truncated SPIKE algorithm [3, 4, 5]. However, we have also derived the estimate

$$\|B^{(j)}\|_\infty \leq (\epsilon^{1+q})^{2^{j-1}}, \quad j = 1, 2, \ldots$$

and this estimate is tight. The estimate is significant, because it characterizes the worst case behavior of incomplete cyclic reduction. This is a work in progress and we expect to have additional results ready before the Householder Symposium in July 2011.

References


On the Best Symmetric Rank-$k$ Approximation of a Symmetric Tensor

*Tamara G. Kolda*

Abstract

Tensors are higher-order generalizations of matrices. We say that a real-valued order-$m$ tensor $\mathbf{A}$ of size $n \times n \times \cdots \times n$ is symmetric if its entries are invariant under any permutation of the indices. Symmetric tensors appear in many applications such as higher-order statistics, blind source signal separation, diffusion weighted tensor imaging, and more.

The goal of the best symmetric rank-$k$ approximation problem is to find real-valued $n$-vectors $\mathbf{u}_1$ through $\mathbf{u}_k$ that minimize

$$
\left\| \mathbf{A} - \sum_{j=1}^{k} \mathbf{u}_j \circ \mathbf{u}_j \circ \cdots \circ \mathbf{u}_j \right\|_2^2 = \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \cdots \sum_{i_m=1}^{n} \left( a_{i_1i_2\ldots i_m} - \sum_{j=1}^{k} u_{i_1j} u_{i_2j} \cdots u_{i_mj} \right)^2.
$$

(1)

I will discuss the challenges of solving this problem, related theoretical work such as [3], some novel algorithms for computing the symmetric rank-$k$ decomposition, and practical applications. My results will build upon recent work in two areas:

- If $k = 1$, then (1) is the best rank-1 approximation problem, which been shown to be equivalent to the problem of finding tensor eigenvalues. In [4], J. Mayo and I introduced and analyzed the Shifted Symmetric Higher-Order Power Method (SS-HOPM) method for computing the best rank-1 approximation of a tensor. In subsequent work [2], G. Ballard and others showed that a GPU implementation can realize a 70X speed-up as compared to a sequential CPU version.

- The nonsymmetric version of (1) corresponds to the CANDECOMP/PARAFAC tensor approximation problem. E. Acar and others have recently developed a direct optimization approach (CP-OPT) [1] that is also amenable to the symmetric rank-$k$ problem.

References


Bivariate Matrix Functions

Daniel Kressner

Abstract

Given a square matrix $A$ and a univariate scalar function $f(z)$ defined on the spectrum of $A$, the matrix function $f(A)$ is again a square matrix of the same size. This talk will be concerned with the following question. What is an appropriate bivariate (multivariate) extension of matrix functions? More specifically, given two square matrices $A$ and $B$ along with a bivariate scalar function $f(x,y)$, is there a sensible way of “evaluating” $f$ at $A$ and $B$?

Definition. The bivariate matrix function $f(A,B)$, $A \in \mathbb{R}^{m \times m}$ and $B \in \mathbb{R}^{n \times n}$, proposed in this talk is not a matrix but a linear operator on $\mathbb{R}^{m \times n}$. For a bivariate polynomial $p(x,y) = \sum_{i=1}^{s} \sum_{j=1}^{t} p_{ij} x^i y^j$, we define

$$p(A,B) : C \mapsto \sum_{i=1}^{s} \sum_{j=1}^{t} p_{ij} A^i B^j.$$

For general $f(x,y)$, we set $f(A,B) = p(A,B)$, where $p$ is a bivariate polynomial interpolating $f$ and sufficiently high mixed derivatives of $f$ at the eigenvalues of $A$ and $B$. In the case of holomorphic $f$,

$$f(A,B)(C) = -\frac{1}{4\pi^2} \oint_{\Gamma_A} \oint_{\Gamma_B} f(x,y)(xI-A)^{-1}(yI-B)^{-T} dy \, dx,$$

where $\Gamma_A, \Gamma_B$ are contours of open sets $\Omega_A, \Omega_B$ containing the eigenvalues of $A$ and $B$, respectively, such that $f$ is holomorphic on $\Omega_A \times \Omega_B$.

Properties. Apart from the Cauchy integral representation (1), bivariate matrix functions have several other convenient properties that resemble well-known properties of the univariate case. For example, the eigenvalues of $f(A,B)$ are given by $f(\lambda,\mu)$, where $\lambda$ is an eigenvalue of $A$ and $\mu$ is an eigenvalue of $B$. More importantly, if $u(z)$ is a univariate function then

$$u(f(A,B)) \equiv (u \circ f)(A,B),$$

meaning that $u$ evaluated (in the usual sense of matrix functions) at the matrix representation of the linear operator $f(A,B)$ gives the matrix representation for the bivariate matrix function $u \circ f$ at $A,B$.

Applications. The concept of bivariate matrix functions unifies linear matrix equations and matrix Fréchet derivatives.

For example, the Sylvester equation $AX + XB^T = C$ can be written as $f_{\text{sylv}}(A,B)(X) = C$ with $f_{\text{sylv}}(x,y) = x + y$. According to (2), its solution can be written as $X = g_{\text{sylv}}(A,B)(C)$ with $g_{\text{sylv}}(x,y) = 1/f_{\text{sylv}}(x,y) = 1/(x + y)$. Similarly, the solution of a Stein equation $X + AXB^T = C$ can be written as $X = g_{\text{stein}}(A,B)(C)$ with $g_{\text{stein}}(x,y) = 1/(1 + xy)$. The Cauchy integral representation (1) and a representation based on the Jordan canonical forms of $A,B$, see [1], generalize results obtained by Krein (1964) and Lancaster (1970) for linear matrix equations.

131
It can be shown that the Fréchet derivatives of a univariate matrix function $f(A)$ in direction $C$ satisfies

$$Df\{A\}(C) = f^{[1]}\{A, A^T\}(C), \quad \text{with} \quad f^{[1]}(x, y) := f[x, y] = \begin{cases} \frac{f(x) - f(y)}{x - y}, & \text{for } x \neq y, \\ f'(x), & \text{for } x = y. \end{cases} \quad (3)$$

As a consequence, existing results for matrix Fréchet derivatives by Horn and Johnson (1991), Mathias (1993), Najfeld and Havel (1995), Mathias (1996), Bhatia and Sinha (1999) immediately follow as special cases from results for bivariate matrix functions. In some cases, the results obtained from the general framework are even stronger. For example, (3) allows us to show the relation

$$f \left( \begin{bmatrix} A & C \\ 0 & A \end{bmatrix} \right) = \begin{bmatrix} A & Df\{A\}(C) \\ 0 & A \end{bmatrix}$$

not only in a very convenient manner but also under weaker conditions on $f$ and $A$ than previously stated.

As a third application, we mention the computation of a univariate matrix function $f(A)$ for a matrix of the form $A = I \otimes A + B \otimes I$, which arises, e.g., from a tensor finite element discretization of a separable partial differential equation. Then $f(A) \doteq \tilde{f}\{A, B\}$ with $\tilde{f}(x, y) = f(x + y)$. For functions like $f(z) = \sqrt{z}$, this relationship may reduce the computational cost for evaluating $f(A)$ significantly.

**Summary and Outlook.** The definition of bivariate matrix function proposed in this talk results in unification and improvements of existing results for linear matrix equations and matrix Fréchet derivatives. It remains to be seen which other applications fit into our framework.

A discussion of computational aspects can be found in [1]. In the talk, we will provide evidence that the concept of bivariate matrix functions also offers a more abstract view and possibly new insights for a variety of other, more advanced results:

1. If $C$ has low rank then it can be shown for quite general situations that the singular values of $f\{A, B\}(C)$ decay rapidly, extending results on the singular value decay of solutions to linear matrix equations.
2. Existing Krylov subspace methods for Lyapunov matrix equations can be extended and viewed as bivariate polynomial matrix approximations.
3. Fréchet derivatives of bivariate matrix functions provide a computationally inexpensive way for studying the sensitivity of linear matrix equations.

The extension of bivariate matrix functions to arbitrary multivariate functions is rather simple and can be found in:

A Critical Path Approach to Analyzing Parallelism of Algorithmic Variants

Julien Langou

Abstract

Dense linear algebra is known for its $O(n^3)$ operations for $O(n^2)$ data movement. Thus, when $n$ is large enough, communication is negligible; efficient parallelism is also achieved (e.g., ScaLAPACK, HPL). However, it is not practically relevant to consider $n$ too large. The size of dense matrices from application scientists remains around at most $10^4$; these dense linear kernels of size $10^2$ to $10^4$ are called often and become bottlenecks for large applications. These kernels need to be accelerated by maximizing parallelism and minimizing communication. However a relevant question is by how much parallelism can be increased and communication can be decreased.

New lower bounds need to be derived to account for these issues. Limitations of our numerical linear algebra algorithms for dense problems are changing. In addition to the computational peak, we now need to take into account the time for communication and the amount of parallelism in our algorithms. A lower bound that was successfully used for over half a century in linear algebra is given by the number of flops to be performed times the peak performance of the computing units. Indeed, it seemed a given that dense linear algebra kernels would always run at peak performance. Current architecture designs (communication cost, hierarchical memory and massive parallelism) and the needs of application scientists ($n$ stays in the 10K) make this lower bound obsolete. Dense linear algebra kernels have become bandwidth-limited, latency-limited, parallelism-limited.

In this talk we focus only on the parallelism part and present a critical path approach to analyzing parallelism of algorithmic variants. (The topic of minimizing communication will be covered in other talks.) We present two case study for our analysis. Firstly Cholesky inversion, secondly QR factorization.

Algorithms come with multiple variants which are obtained by changing the mathematical approach from which the algorithm is derived. These variants offer a wide spectrum of performance when implemented on a parallel computing machine and we seek to understand these differences in performances from a theoretical point of view. To that aim, we derive and present the critical path lengths of each algorithmic variant for our application problem which enables us to determine a lower bound on the time to solution. This metric provides an intuitive grasp of the performance of a variant and we present numerical experiments to validate the tightness of our lower bounds on practical applications.

The first case study is the Cholesky inversion and its use in computing the inverse of a symmetric positive definite matrix. For $t$-by-$t$ tile matrix with $b$-by-$b$ tiles, combining a variety of different approaches, the length of the critical path of our final Cholesky inversion algorithm is $(9t+23)(b^3/3)$ flops. This is to be compared with the length of the critical path of the Cholesky factorization (which represents one third of the work in sequential). Cholesky factorization is $(9t - 10)(b^3/3)$ flops. Computing the Cholesky inverse has asymptotically the same “parallel cost” (within $33(b^3/3)$ flops, a quantity independent of $n$) as computing the Cholesky factorization.

The second case study is the choice of the reduction tree in the factorization of the Communication Avoiding QR. The reduction tree influences the number of communications, the type of communications and the parallelization of the algorithm. Flat tree is best for sequential algorithm and minimizes communication from main memory to computing unit. Binary tree minimizes communication for tall and skinny matrices. Binary tree enables great parallelism for tall and skinny matrices. Flat tree enables great parallelism for square matrices. Depending on the shape of the
matrix and the underlying architectures, we are lead to consider a variety of different trees.

For both of these case studies, we present experimental validation ranging from single node multicore platform to grid computing. We validate that our critical path approach correctly predicts the limitation of the algorithm and that it enables us to select appropriate variants of an algorithm.
Accurate Solutions of $M$-Matrix Algebraic Riccati Equations

Jungong Xue, Shufang Xu, and Ren-Cang Li

Abstract

This talk is concerned with the relative perturbation theory and its entrywise relatively accurate numerical solutions of an $M$-matrix Algebraic Riccati Equations (MARE)

$$XDX - AX - XB + C = 0,$$

by which we mean the following partitioned matrix

$$W = \begin{pmatrix} m & n \\ mB - D & -C \\ n - C & -A \end{pmatrix},$$

is a nonsingular or irreducible singular $M$-matrix. It is known that such an MARE has a unique minimal nonnegative solution $\Phi$:

$$\Phi \leq X \quad \text{for any other nonnegative solution } X \text{ of (1)}.$$

This kind of Riccati equations arise in applied probability and transportation theory and have been extensively studied. One important application is in stochastic fluid models [4, 5], where each entry of the solution has a physical meaning and it is highly desirable and important to investigate the sensitivity of even very tiny entries and to compute them accurately if feasible.

When $D = 0$, MARE (1) degenerates to an $M$-matrix Sylvester equation (MSE)

$$AX + XB = C,$$

which has a unique solution that is nonnegative if $A$ and $B$ are $M$-matrices and one of them is also nonsingular.

Our first goal in this talk is to perform an entrywise perturbation analysis for the minimal nonnegative solution $\Phi$. Specifically, we seek bounds on the entrywise relative errors in the solution:

$$\max_{i,j} |(\tilde{\Phi} - \Phi)_{ij}|/\Phi_{ij}$$

caused by small entrywise relative perturbations to the coefficient matrices $A$, $B$, $C$, and $D$, measured by

$$|\tilde{A} - A| \leq \epsilon|A|, \ |\tilde{B} - B| \leq \epsilon|B|, \ |\tilde{C} - C| \leq \epsilon C, \ |\tilde{D} - D| \leq \epsilon D.$$

Our results in the form of

$$\max_{i,j} |(\tilde{\Phi} - \Phi)_{ij}|/\Phi_{ij} \leq \alpha \epsilon + O(\epsilon^2)$$

suggest each and every entry of the solution, no matter how tiny it may be, are determined to a relative accuracy that is comparable to the entrywise relative accuracy residing in these coefficient matrices, where $\alpha$ is a constant and given explicitly. In a way, the most practically important case is when $W$ is an irreducible singular $M$-matrix. We have results on that case, too.

Previously related results in [1] bound the norm of the solution error: $\|\tilde{\Phi} - \Phi\| \leq \beta\|\tilde{W} - W\|$, where $\beta$ is some constant (not given explicitly). Other results not aiming at MARE in particular may be found in [6], and these results also bound some norm of the error, too.
Following the analysis, we demonstrate that certain fixed point iterations [2], the doubling algorithm [3], and the Newton method [1], all after some minor but crucial implementation changes, can deliver numerical solutions with entrywise relative accuracy as the input data deserve. This contrasts favorably to many other methods most of which are backward stable in the normwise sense and cannot produce solutions with guaranteed entrywise relative accuracy as determined by the input data. This is our second goal.

The material of this talk is based on two recent technical reports of ours on the same topic [7, 8].

References


Towards an Optimal Parallel Approximate Sparse Factorization Algorithm Using Hierarchically Semi-separable Structures

Xiaoye S. Li, Shen Wang, and Jianlin Xia

Abstract

Designing fast algorithms for solving linear systems would have significant impact on large scale scientific and engineering simulations. Much progress has been made in fast algorithms for structured linear systems, such as those involving $\mathcal{H}$-matrices, hierarchically semi-separable (HSS) matrices, and quasiseparable matrices. Nearly linear time factorization algorithms have been developed to solve these systems. A key idea behind these algorithms is to fully exploit numerical low rankness in these structured low-rank matrices.

The complexity of the traditional algorithm for factorizing sparse matrices grows superlinearly in problem size. This was shown in the seminal work by George [2] and Lipton et al. [3]. Specifically, for model PDE problems discretized on regular grids of dimension $k^s$, with $s = 2$ or $3$, and $n = k^s$, when nested dissection ordering is employed, the arithmetic and storage (non-zero entries) complexities are respectively $O(n^{3/2})$ and $O(n \log n)$ for 2D, and $O(n^2)$ and $O(n^{4/3})$ for 3D. These were also shown to be lower bounds if Gaussian elimination is performed faithfully.

Only recently, researchers have discovered that for certain discretized PDEs, throughout sparse Gaussian elimination, the off-diagonal blocks of the intermediate, dense Schur complements (corresponding to the separators in nested dissection) exhibit low rankness. Therefore, these separator submatrices can be represented by a rank structured matrix aforementioned, and the fast structured algorithms can be utilized therein. Blending these rank-structured techniques into sparse Gaussian elimination kernels could break the complexity wall faced by traditional factorization algorithms. The particular approach taken by our group is to modify the traditional multifrontal sparse factorization algorithm [1] by introducing the HSS structures in the frontal and update matrix kernels. This is a convenient way of exploiting low-rank property because the frontal matrices correspond precisely to the separators in the assembly tree from a nested dissection ordering. We have developed a structured multifrontal algorithm and Fortran90 code, where at the bottom levels of the assembly tree we perform partial elimination using the traditional multifrontal method, then at the higher levels we perform the remaining elimination with the HSS structured frontal and update matrices. Our analysis shows that by judiciously choosing the switching level such that the amount of work at the bottom traditional levels is about the same as that at the upper structured levels, both the arithmetic and storage complexities are nearly linear with a prefactor related to the maximum of all HSS ranks of the frontal and update matrices [5].

We now propose a parallel structured multifrontal algorithm that maintains the nearly linear complexity as the sequential one, while at the same time minimizing the amount of communication as much as possible. Below is the outline of our parallel algorithm.

1. Use a parallel nested dissection algorithm (e.g., ParMetis or PT-Scotch) to reorder the matrix and obtain the assembly tree with separator nodes.

2. Perform parallel numerical factorization in two phases with $P$ processors.
   - (a) Lower-level phase: perform classical multifrontal factorization. This is first done in serial at the bottom-most levels with $P$ subtrees, then in parallel at the higher levels when there are fewer than $P$ separators.
   - (b) Upper-level phase: perform parallel structured multifrontal factorization.
• At the switching level, construct HSS representation for the frontal matrices in parallel.
• At each higher level, first eliminate a separator using a parallel HSS factorization algorithm, then perform parallel extend-add to merge the update matrices to the parent.

The success of this approach depends critically on efficient handling of the hierarchical parallelism: the outer coarse-grained parallelism related to the multifrontal factorization which is guided by the assembly tree, and the inner fine-grained parallelism requiring various HSS operations [6]. The outer parallel algorithm has been established by previous research, most notably implemented in the widely used MUMPS software package [4]. Our new challenge is to design the parallel HSS matrix kernels with low communication complexity, and seamlessly integrate them into the outer parallel structure. These kernels include parallel HSS compression via rank revealing QR factorization for frontal matrix representation, parallel HSS QL/LQ factorization for frontal matrix elimination, and parallel data-sparse/structured extend-add for assembling update matrices.

In this talk, we will present parallel performance of the proposed algorithm on leading parallel architectures, will show analysis of both arithmetic and communication complexities, and will comment on the optimality of our parallel algorithm.

References

On the Convergence of GMRES for a Convection-diffusion Model Problem

Jurjen Duintjer Tebbens, Jörg Liesen, and Zdeněk Strakoš

Abstract

A standard model problem in the area of convection-diffusion problems is given by

\[-\nu \Delta u + [0, 1] \cdot \nabla u = f \text{ in } \Omega = (0, 1) \times (0, 1), \quad u = g \text{ on } \partial \Omega,\]

and its discretisation on a uniform \( N \times N \) grid with bilinear finite elements using the streamline upwind Petrov-Galerkin (SUPG) method (sometimes also called the streamline-diffusion finite element (SDFEM) method); see, e.g., [2, Chapter 3]. In short, the idea of the SUPG method is to stabilise the numerical solution by adding an artificial diffusion term in the weak formulation of (1). The amount of artificial diffusion is controlled by a stabilisation parameter \( \delta \geq 0 \). Independent of the value of \( \delta \), the resulting SUPG discretised operator \( A_\delta \) is nonsymmetric.

The GMRES method of Saad and Schultz [6] is frequently used for solving the corresponding linear algebraic system \( A_\delta x = b_\delta \). When looking at the GMRES residual norm curves, one observes an intriguing behavior, which can be summarized as follows:

There typically are two distinct phases of convergence; a slow initial phase followed by a second phase of convergence acceleration.

The length of the initial phase depends on the mesh size \( h = 1/(N + 1) \), i.e., the size of \( A_\delta \), as well as the source term \( f \) and the boundary conditions \( g \) in (1), i.e., the right hand side \( b_\delta \). However, the length is largely independent of all other parameters of the problem.

On the other hand, the speed of the acceleration in the second phase strongly depends on the stabilisation parameter \( \delta \) and its relation to the mesh Peclet number \( P_h = h/(2\nu) \). In particular, the acceleration in the second phase is fastest for a parameter \( \delta^* \) that is “optimal” from the discretisation point of view. This parameter, however, yields a discretised operator \( A_{\delta^*} \) that has very ill-conditioned eigenvectors, and thus can be considered highly nonnormal. This means that the fastest convergence acceleration in the second phase occurs for the “most nonnormal” discretised operator.

Analysing the GMRES convergence behavior for this model problem has been a challenging research topic for many years, which started with the contributions of Fischer et al. [4] and Ernst [3]. An analysis of the initial phase was given in the paper [5], which was presented at Householder XVI in 2005.

This talk will be based on results obtained in [1] that complement the previous work by analyzing the second phase of convergence. In particular, it will be discussed how the acceleration in the second phase is related to the mesh Peclet number \( P_h \) and the choice of the stabilization parameter \( \delta \). The analysis is based on some new expressions and bounds for the GMRES residuals that are generally applicable.

To describe one of these bounds, let a diagonalizable matrix \( A \) of size \( N \times N \) with (at least) \( k + 1 \) distinct eigenvalues \( \sigma_1, \ldots, \sigma_{k+1} \) be given. Let \( A = W \text{diag}(\sigma_1, \ldots, \sigma_N)W^{-1} \) with \( W = [w_1, \ldots, w_N] \) be an eigendecomposition and suppose that the initial residual of the GMRES method is of the form \( r_0 = \sum_{i=1}^{\ell} \theta_i w_i \), where \( \ell \geq k + 1 \) and \( \theta_1, \ldots, \theta_\ell \) are nonzero scalars. Then it can be shown that the \( k \)th GMRES residual norm for \( A \) and \( r_0 \) satisfies

\[ \| r_k \| \geq \frac{\omega_{\min}(W_\ell \Theta_\ell)}{\| e_1^T M_{k+1}^{-1} \|}, \]

where \( \omega_{\min}(W_\ell \Theta_\ell) \) is the minimum generalized eigenvalue of \( W_\ell \Theta_\ell \).
where $\omega_{\text{min}}$ denotes the smallest singular value, $W_\ell \equiv [w_1, \ldots, w_\ell]$, $\Theta_\ell \equiv \text{diag}(\theta_1, \ldots, \theta_\ell)$, and $M_{k+1}$ is the $(k+1) \times (k+1)$ Vandermonde matrix corresponding to the eigenvalues $\sigma_1, \ldots, \sigma_{k+1}$.

The main idea of the approach, in comparison to standard convergence results for GMRES, is that the convergence bounds should reflect the interplay between eigenvalues/eigenvectors and the initial residual. In the lower bound stated above, for example, the initial residual is not separated from the eigenvectors. Using this bound in the context of the convection-diffusion model problem allows to explain why the conditioning of the eigenvectors plays (almost) no role for the speed of acceleration in the second phase. Such bounds can be of interest beyond their application to the convection-diffusion model problem.

References


Fundamental Difficulties of Numerical Multilinear Algebra

Lek-Heng Lim

Abstract

There has been a recent spike of interest in ‘tensor computations’. The idea of developing tensors in numerical computing is appealing. Many problems in computational science and engineering may eventually be reduced to one or more standard problems involving matrices: system of linear equations, least squares problems, eigenvalue problems, singular value problems, low-rank approximations, etc. If one could effectively solve similar problems for tensors of higher order, then one would have substantially enlarged the arsenal of fundamental tools in numerical computations.

We will show that the path from linear to multilinear is nevertheless fraught with fundamental difficulties; these are the consequences of impossibility, nonexistence, and inapproximability results and are unavoidable. We will highlight three specific difficulties in detail and briefly mention several others:

Nonexistence of canonical forms for tensors. There cannot be an equivalent of the Jordan form and the singular value decomposition for tensors nor an equivalent of the eigenvalue decomposition for symmetric tensors. More generally, application of results by Kac, Dadok-Kac, Gabriel, and others, shows that canonical forms for tensors (as well as common subclasses of tensors such as symmetric and alternating ones) cannot exist except in very special cases.

NP-hardness and polynomial-time-inapproximability of tensor problems. The problems of computing rank, best rank-1 approximation, singular value\(^1\), spectral norm of a 3-tensor are all NP-hard; as are the problems of computing eigenvalues\(^1\) of a symmetric 3-tensor, solving a system of bilinear equations, or bilinear least squares. Furthermore, some of these problems are NP-hard to even approximate while others have no fully polynomial time approximation schemes (assuming $\text{NP} \neq \text{ZPP}$).

Nonexistence of optimal low-rank tensor approximations. One of the most common tensor computation problem (often under the heading PARAFAC or CANDECOMP) is that of finding a best rank-$r$ approximation to a given tensor or symmetric tensor. The optima in these optimization problems are often not attainable over a wide range of ranks, dimensions, orders, norms, and could happen on an input set of positive measure. A consequence is that problems such as best rank-$r$ approximations of a tensor, separable approximations of operators by sums of Kronecker products, tensor completion, etc, are unsolvable in general.

Where do these difficulties leave future developments of numerical multilinear algebra? We will argue that the undertaking requires a change in expectations — one should not expect to solve a tensor computation problem in all instances, it is good enough that one solves the problem for a large class of interesting instances. Time permitting, we would also discuss some positive results in this spirit.

\(^1\)That there is no singular value decomposition or eigenvalue decomposition for tensors does not preclude natural definitions of singular values and eigenvalues.
The aim of this work is to devise a reliable algorithm for computing \( A^p \) for \( A \in \mathbb{C}^{n \times n} \) and arbitrary \( p \in \mathbb{R} \). The need to compute fractional powers \( A^p \) arises in a variety of applications, including Markov chain models in finance and healthcare [3], [10], fractional differential equations [9], discrete representations of norms corresponding to finite element discretizations of fractional Sobolev spaces [2], and the computation of geodesic-midpoints in neural networks [5]. Here, \( p \) is an arbitrary real number, not necessarily rational. Often, \( p \) is the reciprocal of a positive integer \( q \), in which case \( X = A^p = A^{1/q} \) is a \( q \)th root of \( A \). Various methods are available for the \( q \)th root problem, based on the Schur decomposition and appropriate recurrences, Newton or inverse Newton iterations, Padé iterations, or a variety of other techniques; see [6, Chap. 7] and [7] for surveys. However, none of these methods is applicable for arbitrary real \( p \).

Arbitrary matrix powers can be defined via the Cauchy integral, which yields many different matrices \( A^p \), as the branch of the function \( z^p \) can be chosen independently around each eigenvalue. For practical purposes it is more useful to define \( A^p \) uniquely as follows.

**Definition 1.** For \( A \in \mathbb{C}^{n \times n} \) with no eigenvalues on the close negative real axis and \( p \in \mathbb{R} \),

\[
A^p = \exp (p \log(A)),
\]

where \( \log(A) \) is the principal logarithm of \( A \) [6, Thm. 1.31].

Note that for \( p = 1/q \), with \( q \) a positive integer, \( A^p \) from (1) reduces to the principal \( q \)th root of \( A \) [6, Thm. 7.2].

When \( A \) is diagonalizable, so that \( A = XDX^{-1} \) for a diagonal \( D = \text{diag}(d_i) \) and nonsingular \( X \), we can compute \( A^p = XD^pX^{-1} = X\text{diag}(d_i^p)X^{-1} \). Alternatively, for any \( A \) we can compute the Schur decomposition \( A = QTQ^* \), with \( Q \) unitary and \( T \) upper triangular, from which \( A^p = QT^pQ^* \). The matrix \( T^p \) has diagonal elements \( t_{ii}^p \) and we can obtain the superdiagonal elements from the Parlett recurrence if the \( t_{ii} \) are distinct [6, sec. 4.6]. However, this approach breaks down when \( A \) is nonnormal with repeated eigenvalues.

The definition (1) suggests another way to compute \( A^p \): to employ existing algorithms for the matrix exponential and the matrix logarithm. However, if we use the inverse scaling and squaring method for \( X = \log(A) \) [4] followed by the scaling and squaring method for \( \exp(pX) \) [1] then we are computing two Padé approximants: one of the logarithm and the other of the exponential. We expect benefits to accrue from employing a single Padé approximant, to \((1 - x)^p\).

In this work a new algorithm is developed for computing arbitrary real powers \( A^p \) of a matrix \( A \in \mathbb{C}^{n \times n} \). Our algorithm exploits the relation \( A^p = (A^{1/2k})^{p2^k} \). In order to facilitate the computation of the square roots we compute an initial Schur decomposition \( A = QTQ^* \), so that the problem is reduced to that for a triangular matrix. The new algorithm takes \( k \) square roots of the triangular factor \( T \), evaluates an \([m/m]\) Padé approximant of \((1 - x)^p\) at \( I - T^{1/2k} \), and squares the result \( k \) times. This approach is analogous to the inverse scaling and squaring method for the matrix logarithm [4]. The parameters \( k \) and \( m \) are chosen to minimize the cost subject to achieving double precision accuracy in the evaluation of the Padé approximant, making use of a result that bounds the error in the matrix Padé approximant by the error in the scalar Padé approximant.

A Schur–Padé Algorithm for Fractional Powers of a Matrix

Nicholas J. Higham, Lijing Lin

Abstract

The aim of this work is to devise a reliable algorithm for computing \( A^p \) for \( A \in \mathbb{C}^{n \times n} \) and arbitrary \( p \in \mathbb{R} \). The need to compute fractional powers \( A^p \) arises in a variety of applications, including Markov chain models in finance and healthcare [3], [10], fractional differential equations [9], discrete representations of norms corresponding to finite element discretizations of fractional Sobolev spaces [2], and the computation of geodesic-midpoints in neural networks [5]. Here, \( p \) is an arbitrary real number, not necessarily rational. Often, \( p \) is the reciprocal of a positive integer \( q \), in which case \( X = A^p = A^{1/q} \) is a \( q \)th root of \( A \). Various methods are available for the \( q \)th root problem, based on the Schur decomposition and appropriate recurrences, Newton or inverse Newton iterations, Padé iterations, or a variety of other techniques; see [6, Chap. 7] and [7] for surveys. However, none of these methods is applicable for arbitrary real \( p \).

Arbitrary matrix powers can be defined via the Cauchy integral, which yields many different matrices \( A^p \), as the branch of the function \( z^p \) can be chosen independently around each eigenvalue. For practical purposes it is more useful to define \( A^p \) uniquely as follows.

**Definition 1.** For \( A \in \mathbb{C}^{n \times n} \) with no eigenvalues on the close negative real axis and \( p \in \mathbb{R} \),

\[
A^p = \exp (p \log(A)),
\]

where \( \log(A) \) is the principal logarithm of \( A \) [6, Thm. 1.31].

Note that for \( p = 1/q \), with \( q \) a positive integer, \( A^p \) from (1) reduces to the principal \( q \)th root of \( A \) [6, Thm. 7.2].

When \( A \) is diagonalizable, so that \( A = XDX^{-1} \) for a diagonal \( D = \text{diag}(d_i) \) and nonsingular \( X \), we can compute \( A^p = XD^pX^{-1} = X\text{diag}(d_i^p)X^{-1} \). Alternatively, for any \( A \) we can compute the Schur decomposition \( A = QTQ^* \), with \( Q \) unitary and \( T \) upper triangular, from which \( A^p = QT^pQ^* \). The matrix \( T^p \) has diagonal elements \( t_{ii}^p \) and we can obtain the superdiagonal elements from the Parlett recurrence if the \( t_{ii} \) are distinct [6, sec. 4.6]. However, this approach breaks down when \( A \) is nonnormal with repeated eigenvalues.

The definition (1) suggests another way to compute \( A^p \): to employ existing algorithms for the matrix exponential and the matrix logarithm. However, if we use the inverse scaling and squaring method for \( X = \log(A) \) [4] followed by the scaling and squaring method for \( \exp(pX) \) [1] then we are computing two Padé approximants: one of the logarithm and the other of the exponential. We expect benefits to accrue from employing a single Padé approximant, to \((1 - x)^p\).

In this work a new algorithm is developed for computing arbitrary real powers \( A^p \) of a matrix \( A \in \mathbb{C}^{n \times n} \). Our algorithm exploits the relation \( A^p = (A^{1/2k})^{p2^k} \). In order to facilitate the computation of the square roots we compute an initial Schur decomposition \( A = QTQ^* \), so that the problem is reduced to that for a triangular matrix. The new algorithm takes \( k \) square roots of the triangular factor \( T \), evaluates an \([m/m]\) Padé approximant of \((1 - x)^p\) at \( I - T^{1/2k} \), and squares the result \( k \) times. This approach is analogous to the inverse scaling and squaring method for the matrix logarithm [4]. The parameters \( k \) and \( m \) are chosen to minimize the cost subject to achieving double precision accuracy in the evaluation of the Padé approximant, making use of a result that bounds the error in the matrix Padé approximant by the error in the scalar Padé approximant.
with argument the norm of the matrix. The Padé approximant is evaluated from the continued fraction representation in bottom-up fashion, which is shown to be numerically stable. In the squaring phase the diagonal and first superdiagonal are computed from explicit formulae for $T^{p/2}$, yielding increased accuracy. Since the basic algorithm is designed for $p \in (-1, 1)$, a criterion for reducing an arbitrary real $p$ to this range is developed, making use of bounds for the condition number of the $A^p$ problem. How best to compute $A^k$ for a negative integer $k$ is also investigated. In numerical experiments the new algorithm is found to be superior in accuracy and stability to several alternatives, including the use of an eigendecomposition and approaches based on separate approximation of the exponential and logarithm in the formula $A^p = \exp(p \log(A))$ using the best available methods. It is also superior to the current MATLAB implementation of $A^{-p}$.

References

A Numerical Linear Algebra View of the Tao-Vu Smallest Singular Value Limit and the SDO Extension

Alan Edelman and Po-Ru Loh

Abstract

We revisit the limiting distribution of the smallest singular value $\sigma_n$ of an $n \times n$ random matrix with iid mean zero, variance 1 entries. Extensive computations performed in 1989 left little doubt that as $n$ approaches infinity, $\sqrt{n} \sigma_n$ approaches a limiting distribution independent of the distribution from which the entries are drawn. Twenty years later, Tao and Vu proved this conjecture (under a finite moment assumption) using an approach motivated by "property testing" from theoretical computer science.

We will dissect the result from a numerical analyst’s perspective, starting with a reformulation of the intuition in terms of sampling a submatrix from the QR decomposition. We then explore data comparing the strength of bounds and accuracy of estimators to reality – which we hope will pique the interest of the experimentally-minded – and also argue a link (via bidiagonalization) to an underlying stochastic differential operator – an alternative approach which should interest the theoretically-oriented. This work is joint with Alan Edelman.
Möbius Transformations of Matrix Polynomials

D. Steven Mackey, Niloufer Mackey, Christian Mehl, and Volker Mehrmann

Abstract
We consider \( n \times n \) matrix polynomials of the form \( P(\lambda) = \sum_{i=0}^{k} \lambda^i A_i \), where \( A_k \in \mathbb{F}^{n \times n}, A_k \neq 0 \). Here \( \mathbb{F} \) denotes the field \( \mathbb{R} \) or \( \mathbb{C} \). The numerical solution of the associated polynomial eigenvalue problem \( P(\lambda)x = 0 \) is one of the most important tasks in the vibration analysis of buildings, machines, and vehicles. In many applications, the coefficient matrices have a further structure which reflects the properties of the underlying physical model. Examples of such structure include palindromic, odd, even, and Hermitian. The algebraic relations between the coefficient matrices \( P(\lambda) \) lead to symmetries in the eigenvalues and eigenvectors of \( P \).

It is well known that the Cayley transformation and its generalization to matrix pencils relates Hamiltonian structure to symplectic structure for both matrices and pencils.

In earlier work we extended [1] the classical definition of Cayley transformation to matrix polynomials. Particular Cayley transformations \( C_{-1}(P) \) and \( C_{+1}(P) \) were used to connect palindromic and even/odd matrix polynomials and their variants obtained using transpose or conjugate-transpose. These transformations are given by

\[
C_{-1}(P)(\mu) := (\mu + 1)^k P \left( \frac{\mu - 1}{\mu + 1} \right), \quad C_{+1}(P)(\mu) := (1 - \mu)^k P \left( \frac{1 + \mu}{1 - \mu} \right). \tag{1}
\]

Since the eigenvalues of palindromic matrix polynomials occur in reciprocal pairs, and those of even/odd polynomials in plus-minus pairs, our attention was restricted to fractional linear transformations of the complex plane that map reciprocal pairs \( (\mu, 1/\mu) \) to plus/minus pairs \( (\lambda, -\lambda) \).

We now consider general fractional linear transformations, \( \lambda \mapsto \frac{a\lambda + b}{c\lambda + d} \). The matrix polynomial obtained mutatis mutandis as in (1), is called the Möbius transformation of \( P \).

We systematize and unify previously known results, and develop new properties of this very useful device that have wider significance. They apply to general matrix polynomials, regular and singular, square and rectangular. We track the effect on Jordan characteristic and partial multiplicities of \( P \), and discuss the implications for strong and weak linearizations of \( P \).

References

The Quadratic Realizability Problem for Matrix Polynomials

D. Steven Mackey, with F. De Terán and F. Dopico, Maha Al-Ammari and Francoise Tisseur

Abstract

Recent years have seen intense investigation of the polynomial eigenproblem $P(\lambda)x = 0$, where $P(\lambda) = \sum_{j=0}^{k} \lambda^j A_j$ is an $n \times n$ matrix polynomial. This problem arises in many applications, including structural dynamics, vibrational analysis, control systems, and differential-algebraic equations, to give a few examples. One commonly used approach to polynomial eigenproblems is via linearizations. A linearization of an $n \times n$ matrix polynomial $P(\lambda)$ of degree $k$ is a $kn \times kn$ pencil $L(\lambda) = \lambda X + Y$ such that $E(\lambda)L(\lambda)F(\lambda) = \text{diag}(P(\lambda), I_{k(n-1)})$ for some unimodular $E(\lambda)$ and $F(\lambda)$. Consequently $L(\lambda)$ has the same set of eigenvalues and the same elementary divisor structure as $P(\lambda)$. The use of such pencils $L(\lambda)$ is now well-established as a theoretical and computational tool for studying the properties of matrix polynomials [2, 6].

In many applications the underlying polynomial $P(\lambda)$ has additional algebraic structure, e.g., palindromic, alternating, or Hermitian structure. For such polynomials it can be advantageous to employ a linearization $L(\lambda)$ with the same algebraic structure as $P$ [4, 7]. If a structure-preserving algorithm is then used [5, 8] to compute the eigenvalues of $L(\lambda)$, the spectral symmetries of the original polynomial $P(\lambda)$ can be guaranteed to be present in the computed eigenvalues, despite the errors inherent to numerical computation with floating-point numbers. This is the structured linearization strategy for structured polynomial eigenproblems. What is the scope of this structured linearization strategy? Recent work has probed this issue from several angles: by constructing large families of structured linearizations [4, 7], analyzing their conditioning and backward error properties [1, 3], and by investigating the possible Smith forms of important classes of structured polynomials [9, 10].

A characterization of the Smith forms of a class $S$ of structured polynomials is an essential prerequisite to being able to say exactly which $P(\lambda) \in S$ have a linearization in $S$ and which do not, thus bounding the scope of the structured linearization strategy for the class $S$. Furthermore, whenever a structured linearization is known to exist in principle, it is important to have some simple way to explicitly construct a structured linearization directly from the matrix coefficients of $P$. Progress towards these ends was made in [9, 10], but an unexpected phenomenon has come into focus as a result: a fundamental dichotomy between the behavior of structured matrix polynomials of even degree vs. those of odd degree. Odd degree palindromic (or alternating) polynomials all have the same constraints on their elementary divisors, as do all even degree palindromic (or alternating) polynomials, but these conditions are different for even and odd degree. As a consequence it can be shown that there exist even degree structured polynomials that have no structured linearizations at all. This dichotomy also strongly suggests that every odd degree palindromic (or alternating) polynomial will have a structured linearization. An explicit example of such a structured linearization for any odd degree palindromic (or alternating) polynomial is presented in each of [9, 10], but what about even degree structured polynomials? Although structured linearizations do not always exist, the Smith form results in [9, 10] strongly suggest that every even degree structured polynomial $P(\lambda)$ may have a structured quadratification, i.e., a quadratic polynomial $Q(\lambda)$ with the same algebraic structure as $P$, and exactly the same elementary divisors as $P$. Investigating this question leads inexorably to the quadratic realizability problem (QRP) for matrix polynomials:

Characterize those lists $L$ of elementary divisors (allowing both finite and infinite elementary divisors in $L$, with repetitions) that comprise the complete list of elementary divisors of some regular quadratic matrix polynomial. For each such quadratically realizable list
show how to concretely construct a specific regular quadratic matrix polynomial that realizes the list $\mathcal{L}$.

A further motivation for this line of investigation comes from the growing number of algorithms currently being developed that are specifically designed to act directly on quadratic matrix polynomials, without the intervention of any linearization.

Structured versions of the QRP can also be formulated. Given a class $\mathcal{S}$ of structured matrix polynomials, characterize (and constructively realize) those lists of elementary divisors that comprise the complete list of elementary divisors of some quadratic matrix polynomial in $\mathcal{S}$. Thus we have the palindromic QRP, the alternating QRP, the Hermitian QRP, ..., etc. Clearly the QRP is a kind of inverse problem, but a somewhat restricted one. Only elementary divisor structure is considered here, not eigenvectors or Jordan chains or sign characteristic, nor anything about any singular structure, comprised of minimal indices and minimal bases.

In this talk I will first describe the surprisingly-simple-to-state (but not-so-easy-to-prove) solution of the QRP for general matrix polynomials. Then I will discuss how this result can be modified to accommodate various structure classes of matrix polynomials, and thus give complete solutions to the palindromic, alternating, and Hermitian QRPs. Finally I will discuss the impact of these results on the existence of quadratifications of even degree matrix polynomials, both with and without structure, as well as giving explicit constructions of some new quadratic companion forms.

References


147
QPBLUR: A Regularized Active-set Method for Sparse Convex Quadratic Programming

Christopher Maes and Michael Saunders

Abstract

We develop an active-set algorithm for convex quadratic programs (QPs) that have many degrees of freedom:

$$\text{minimize } \quad c^T x + \frac{1}{2} x^T H x$$
subject to \quad $$\ell \leq (x, Ax) \leq u,$$

where $A$ and $H$ are sparse, $H$ is positive definite or semidefinite, and the number of variables may be far more than the number of active constraints. The algorithm employs primal and dual regularization within what we call a BCL method (bound-constrained augmented Lagrangian method). This leads to a sequence of QP subproblems that are strictly convex and feasible, and whose KKT systems are guaranteed to be nonsingular for any active set.

A simplified, single-phase algorithm is possible for each QP subproblem. There is no need to control the inertia of the KKT system defining each search direction, and a simple step-length procedure may be used without risk of cycling in the presence of degeneracy. Since all KKT systems are nonsingular, they can be factored with a variety of sparse direct linear solvers such as UMFPACK or MA57 (and no basis repair or KKT repair is needed). Block-LU updates allow for active-set changes without altering the initial KKT factors.

The principal benefit of primal and dual regularization is that warm starts are possible from any given active set. This is vital inside a sequential quadratic programming (SQP) method for nonlinear optimization such as SNOPT. The algorithm has been implemented as the QPBLUR solver (MATLAB and Fortran 95 versions), and the Fortran version has been integrated into SNOPT. We evaluate the performance of QPBLUR on a test set of large convex QPs, and on the sequences of QPs arising from SNOPT’s SQP method.

QPBLUR provides a reliable approach for solving sparse examples of Paige’s generalized least-squares problem. Ordinary least-squares problems with Tikhonov regularization and bounds can be solved as a single QP subproblem.
Fast Approximation of Matrix Coherence

Michael W. Mahoney

Abstract

The concept of matrix coherence measures the extent to which the singular vectors of a matrix are correlated with the standard basis, and as such it is of interest in the recently-popular problem of matrix completion. A more refined concept is that of statistical leverage. Statistical leverage is usually quantified by the diagonal elements of the projection matrix onto the best rank-\(k\) approximation to a matrix. Thus, it is useful in large-scale diagnostic data analysis for identifying outliers in large data matrices and data graphs. Moreover, it is the key structural nonuniformity that must be dealt with (i.e., either rapidly approximated or rapidly uniformized at the preprocessing step) in developing fast random sampling and random projection algorithms for matrix problems such as least-squares regression and low-rank matrix approximation.

As a corollary of our main result, we obtain an algorithm to approximate the coherence of an arbitrary matrix in time qualitatively faster than the naive algorithm. Our main result is a randomized algorithm that takes as input an arbitrary \(m \times n\) matrix \(A\) and a rank parameter \(k\); it returns as output a relative-error approximation to every diagonal element of the projection matrix onto the best rank-\(k\) approximation to \(A\); and it runs in time qualitatively faster than the time to compute a basis for that space. In particular, given an \(m \times n\) matrix \(A\), with \(m \gg n\), the algorithm returns relative-error approximations to all the diagonal elements of the projection matrix onto the left singular subspace in roughly \(O(mn \log n)\) time, as opposed to \(O(mn^2)\) time required by the naive algorithm. In addition, numerical implementations run faster than traditional deterministic algorithms for matrices as small as hundreds by thousands. Our analysis boils down to computing a relative-error approximation to an underconstrained least-squares approximation, or relatedly it can be viewed as an application of Johnson-Lindenstrauss ideas.
Low Displacement Rank Representations for the Spectral Factorization of Matrix Polynomials

Alexander Malyshev

Abstract

Given a matrix polynomial

\[ p(\lambda) = t_0 + t_1 \lambda + \cdots + t_m \lambda^m, \quad t_i \in \mathbb{C}^{n \times n}, \]

which has no eigenvalues on the unit circle, the block Toeplitz matrices

\[
A_0 = \begin{bmatrix}
t_0 & t_1 & \cdots & t_{m-1} \\
t_0 & t_1 & \cdots & t_{m-2} \\
\vdots & \ddots & \ddots & \vdots \\
t_0 & & & t_0 \\
\end{bmatrix}, \quad B_0 = \begin{bmatrix}
t_m & t_{m-1} & t_m \\
t_{m-1} & t_m & \ddots & \vdots \\
\vdots & \ddots & \ddots & t_1 \\
t_1 & \cdots & t_{m-1} & t_m \\
\end{bmatrix}
\]

determine the linear matrix pencil \( zB_0 - A_0 \), which has no eigenvalues on the unit circle. The matrix iteration

\[
A_{j+1} = A_j(A_j + B_j)^{-1}A_j, \quad B_{j+1} = B_j(A_j + B_j)^{-1}B_j, \quad j = 0, 1, \ldots,
\]

quadratically converges to \( A_\infty \) and \( B_\infty \). The sequence of matrices \( P_j = (A_j + B_j)^{-1}B_j \) quadratically converges to the spectral projector \( P = (A_\infty + B_\infty)^{-1}B_\infty \) onto the right deflating subspace of \( zB_0 - A_0 \) corresponding to the eigenvalues inside the unit circle. The iteration (1) is an analog of the matrix sign function iteration.

If the displacement operator \( \nabla(M) = ZM -MZ \) is determined by the block shift matrix

\[
Z = \begin{bmatrix}
0 & & & \\
I & 0 & & \\
& \ddots & \ddots & \\
& & I & 0
\end{bmatrix},
\]

and \( \nabla(M) = \sum_{k=1}^{r} u_k v_k^T \), where \( u_k \) and \( v_k \) are block vectors, then

\[
M = \mathcal{L}(Me_1) - \sum_{k=1}^{r} \mathcal{L}(u_k)[\mathcal{L}(Zv_k)]^T,
\]

where \( e_1 = [I \ 0 \ \cdots \ 0]^T \) and \( \mathcal{L}(u) \) is the block low triangular block Toeplitz matrix whose the first block column is a block vector \( u \). When \( r \) is small, the matrix \( M \) is called a matrix of a low displacement rank. The Toeplitz-like structure (2) with the low displacement rank is exploited in the so called fast and superfast algorithms for computation of \( Mw \) or \( M^{-1}w \), where \( w \) is an arbitrary vector, see, e.g., [3].

The matrices from the iteration (1) are of a low displacement rank:

\[
\text{rank}\nabla(A_j) = \text{rank}\nabla(B_j) = \text{rank}\nabla(A_j + B_j) = \text{rank}\nabla(P_j) = 2n.
\]

Low displacement rank expressions for \( A_j, \ B_j, \ A_j + B_j \) and \( P_j \) are computed by the following algorithm.
Let us introduce the matrices
\[ L_1 = \mathcal{L}( \begin{bmatrix} \ t_m \\ \vdots \\ t_2 \\ t_1 \end{bmatrix} ), \quad L_2 = \mathcal{L}( \begin{bmatrix} \ t_0 & t_1 & \ldots & t_{m-1} \end{bmatrix}^T ), \quad e_m = [0 \ldots 0 \ I ]^T, \]
and compute two sequences of block vectors
\[ u_0 = e_1, \quad u_{j+1} = A_j(A_j + B_j)^{-1}u_j, \quad j = 0, 1, \ldots, \]
\[ v_0 = e_m, \quad v_{j+1} = A_j^T(A_j^T + B_j^T)^{-1}v_j, \quad j = 0, 1, \ldots. \]
Then
\[ A_j = L_1[\mathcal{L}(Zv_j)]^T - \mathcal{L}(u_j)L_2^T, \]
\[ A_j + B_j = L_1[\mathcal{L}(Z\{2v_j\} + e_1)]^T - \mathcal{L}(2u_j - e_1)L_2^T, \]
\[ P_j = [\mathcal{L}(Z\{2v_{j+1} - v_j - A_j^T(A_j^T + B_j^T)^{-1}e_m\} + e_1)]^T + \mathcal{L}(A_j + B_j)^{-1}u_j)L_2^T. \]

The above representations were inspired by [1].

Assume that \( p(\lambda) = q(\lambda)d(\lambda) \), where all eigenvalues of the monic divisor
\[ d(\lambda) = d_0 + d_1\lambda + \cdots + d_{l-1}\lambda^{l-1} + I\lambda^l \]
lie inside the unit circle, and all eigenvalues of the matrix polynomial \( q(\lambda) \) of degree \( m-l \) lie outside the unit circle. Such factorization is referred to as the spectral one.

Let the columns of a matrix \( V \in C^{mn \times ln} \) span the right deflating subspace of the pencil \( zB_0 - A_0 \) corresponding to the eigenvalues inside the unit circle that is the image subspace of \( P \). Existence of the spectral factorization implies nonsingularity of the square block \( V_d \in C^{ln \times ln} \) composed of the first \( ln \) rows of \( V \), see, e.g., [2].

The coefficients of the spectral divisor \( d(\lambda) \) are given by the formula
\[ \begin{bmatrix} \ d_0 & d_1 & \ldots & d_{l-1} \end{bmatrix} = -V_{d+1}V_d^{-1}, \]
where \( V_{d+1} \in C^{n \times ln} \) denotes the block of rows \( ln + 1, ln + 2, \ldots, ln + n \) of \( V \).

A straightforward extraction of blocks \( V_d \) and \( V_{d+1} \) from \( P_j \) (for sufficiently large \( j \)) is as follows: first compute the QR factorization \( QR = [\mathcal{L}(Z\{2v_{j+1} - v_j - A_j^T(A_j^T + B_j^T)^{-1}e_m\} + e_1), \ L_2]^T \), then the LQ factorization \( LQ_1 = [I, \ \mathcal{L}((A_j + B_j)^{-1}u_j)]Q \). As a result, \( V_d = L(1: ln, 1: ln) \) and \( V_{d+1} = L(ln + 1: ln, n + 1: ln) \).

References


Nonlinearly Structured Low-Rank Approximation
with Application to Algebraic Curve Fitting

Ivan Markovsky

Abstract

Background

The low-rank approximation problem:

\[ \text{Given a matrix } A \in \mathbb{R}^{m \times n}, \text{ an integer } r < \min(m, n), \text{ and a matrix norm } \| \cdot \|, \]
\[ \text{minimize over } B \in \mathbb{R}^{m \times n} \text{ with rank}(B) \leq r, \text{ the approximation error } \| A - B \|. \]

is a prototypical data modeling problem [5, 7]. For applications in systems, control, and signal processing, however, a modification of the basic problem defined above is required. In the context of linear time-invariant models, the matrix \( A \) is Hankel structured and its approximation \( B \) is required to have the same structure. While in some special cases (e.g., Frobenius and spectral norm \( \| \cdot \| \)) the basic low-rank approximation problem can be solved globally by the singular value decomposition, in the presence of structure constraints, global optimality results are not available.

Existing methods for structured low-rank approximation split into two broad classes: convex relaxations and local optimization. Convex relaxations use subspace type methods [10] and the recently developed methods, based on the nuclear norm heuristic [1]. Local optimization type methods are subdivided into methods based on the variable projections [3] and methods based on the alternating projections.

Novelty

A matrix structure is a function from a parameter space to a set of matrices. The Hankel structure and its relatives—block Hankel, block Hankel-Hankel block, Toeplitz, Sylvester, etc.—are linear structures. Existing methods for structured low-rank approximation essentially exploit this fact. In [8] and [4] nonlinearly structured low-rank approximation is considered. An equivalent to the considered in [4] Vandermonde structured low-rank approximation problem, however, is a linearly (Hankel) structured low-rank approximation problem, so that the nonlinear structure can be avoided in this particular case.

The novelty of the present work is in establishing an equivalence between the problem of geometric fitting of algebraic curves to data and the problem of polynomially structured low-rank approximation. We give conditions on the data and the model class that ensure existence and uniqueness of an exact solution and develop methods, algorithms, and software for polynomially structured low-rank approximation. Numerical results showing the application of the developed methods to algebraic curve fitting, are available at http://users.ecs.soton.ac.uk/im/curve-fitting.pdf

The nonlinear low-rank approximation framework leads to conceptual unification of curve fitting methods as well as a practical benefit of using a single algorithm and a piece of software for solving a wide range of curve fitting problems. In particular, the low-rank approximation setting reveals that the class of algebraic fitting methods [6] are relaxations, obtained by removing the structure constraint, of corresponding geometric fitting problems [2].
Related work

Our approach originates from the systems and control community, where of main interest is modeling of linear time-invariant dynamical systems (system identification). The results, however, has links to problems and methods in computer vision, machine learning, computer algebra, and numerical linear algebra.

- Existence and uniqueness of an exact solution of the algebraic curve fitting problem is a special multivariable interpolation problem, where the interpolation points are on an algebraic curve and of interest is recovery of the curve rather than its representation (the multivariable polynomial).
- There is a large body of work on ellipsoid fitting (see, e.g., [2, 6] and the references there in), which is a special case of the considered data fitting problem when the total degree of the polynomial is two.
- The convex relaxation methods for nonlinearly structured low-rank approximation, based on ignoring the nonlinear structure and thus solving the problem as unstructured low-rank approximation, is known in the machine learning literature as kernel principal component analysis [9].
- The principal curves and other dimensionality reduction methods [11] are formulated as fitting an affine variety to data but are solved by alternative methods.

References

An Algorithm for Computing and Updating a New Factorization of Large Symmetric Indefinite Matrices

Nicola Mastronardi, and Paul Van Dooren

Abstract

Indefinite symmetric matrices occur in many applications, such as optimization, partial differential equations and variational problems where they are linked to a so-called saddle point problem. In these applications one is often interested in computing an estimate of the dominant eigenspace of such matrices, in order to solve regularized least squares problems or compute preconditioners. In this talk we propose an incremental method to compute the $UTU^T$ factorization of a symmetric indefinite matrix, where $U$ is an orthogonal matrix and $T$ is a symmetric anti-triangular one, i.e., a matrix having zero entries below the anti-diagonal. Moreover, we describe an algorithm for computing an estimate of the dominant eigenbasis of such matrices based on low rank updates and downdates of indefinite matrices.

We show that the proposed algorithms are well-suited for large scale problems since they are efficient in terms of complexity as well as data management.

Some numerical experiments showing the behavior of the proposed algorithms are presented.

References


The Solution of a Two-Parameter Eigenvalue Problem and a Connection with the Implicitly Restarted Arnoldi Method

Karl Meerbergen

Abstract

Consider the eigenvalue problem

\[(A + \lambda B)x = \mu Mx\]  \hspace{1cm} (1)

where \(\mu\) is the eigenvalue, \(\lambda\) is a (real) parameter and \(A, B,\) and \(M\) are real large square matrices. The goal is to find \(\lambda\) for which (1) has purely imaginary \(\mu\). Such problems arise in the study of the stability of steady state solutions of dynamical systems. Mathematically, this problem can be written as finding the solutions of the eigenvalue problem

\[(A \otimes M + M \otimes A)u + \lambda (B \otimes M + M \otimes B)u = 0 . \]  \hspace{1cm} (2)

The problem can be written as the matrix equation problem

\[AUM^T + MU^T + \lambda (BUM^T + MUB^T) = 0 \]  \hspace{1cm} (3)

which we call the Lyapunov eigenvalue problem. The \(n \times n\) matrix \(U\) is the ‘eigenvector’. If, for a given \(\lambda\), \(\pm \mu\) are the only purely imaginary eigenvalues of (1) and are simple, then \(\lambda\) is a double eigenvalue of (2). Then there is a unique (up to a scalar factor) real symmetric ‘eigenvector’ \(U\) associated with \(\lambda\), which appears to have rank two. This unique matrix can be found by an inverse iteration method on (3) [3]. Imposing symmetry allows to use large scale Lyapunov solvers. The difficulty, however, is that inverse iteration on (3) is infeasible since the iterates are full rank dense matrices. When the method starts to converge, \(U\) can be well approximated by a rank two matrix. In the first iterations, however, this is usually not possible, which makes the method expensive. We therefore introduce a projection step that has two important consequences: the iterates now have rank two since they arise from the solution of a small eigenvalue problem of the form (3); and convergence is faster than inverse iteration.

Equation (1) can be solved by the two-parameter Jacobi-Davidson method [1], but this requires good starting values for \(\lambda\) and \(\mu\). Where, usually, starting values for \(\lambda\) are available, this is often not the case for \(\mu\). One advantage of our method is that a starting value of \(\mu\) is not required. As a result, our method can be used for finding starting values for [1].

In this talk, we make the somehow straightforward extension to inverse subspace iteration. This extension is important to make the connection with the implicitly restarted Arnoldi method. For the special case \(B = M\), and when the Arnoldi method is used as Lyapunov solver [6], we can show a nice interpretation that may help understand the inverse iteration method for (3). In this case, the method is the implicitly restarted Arnoldi method applied to \(A^{-1}M\) with a proper selection of exact shifts and one zero shift [4]. In that respect, the proposed method can be viewed as a combination of subspace iteration and projection as the implicitly restarted Arnoldi method [2] and has similar convergence properties.

We finally will present an extension of the method to the solution of eigenvalue problems with a nonlinear parameter \(\lambda\), as in

\[A(\lambda)x = \mu Mx .\]
We discuss several ideas that can be used to solve the nonlinear problem using inverse residual iteration and Jacobi-Davidson [5] for the nonlinear Kronecker eigenvalue problem
\[(A(\lambda) \otimes M + M \otimes A(\lambda))u = 0 .\]

We show numerical examples to illustrate the methods.

References


The Canonical Generalized Polar Decomposition

Nicholas J. Higham, Christian Mehl, and Françoise Tisseur

Abstract

The polar decomposition is a much-studied matrix decomposition, from the points of view of both theory and computation. Depending on the matrix dimensions and the uniqueness required of the factors, there are several possibilities to define the polar decomposition. If \( A \in F^{m \times n} \) (here \( F \) denotes either the field \( \mathbb{R} \) of real numbers or the field \( \mathbb{C} \) of complex numbers), then the polar decomposition of \( A \) is a decomposition of the form

\[
A = U H,
\]

where \( U \in F^{m \times n} \) is a matrix with orthonormal columns and \( H \in F^{n \times n} \) is a Hermitian positive semidefinite matrix. In this version, the factor \( H \) is uniquely determined (in fact \( H = (A^* A)^{1/2} \)), whereas the factor \( U \) is unique if and only if \( A \) has full rank.

In order to enforce uniqueness of the factor \( U \) as well, one can consider decompositions of the form (1), where \( H \in F^{n \times n} \) is Hermitian positive semidefinite and \( U \in F^{m \times n} \) is a partial isometry, i.e., it satisfies \( \|U x\|_2 = \|x\|_2 \) for all \( x \in \text{range}(U^*) \), or equivalently, \( U^+ = U^* \), where \( U^+ \) denotes the Moore-Penrose pseudoinverse. The so-called canonical polar decomposition of \( A \) is then defined to be such a decomposition, where the partial isometry \( U \) satisfies in addition that \( \text{range}(U^*) = \text{range}(H) \).

The factors \( H \) and \( U \) are then given by \( H = (A^* A)^{1/2} \) and \( U = A H^+ \). In particular, if

\[
A = P \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} Q^*
\]

is a singular decomposition of \( A \), where \( r = \text{rank}(A) \), then

\[
U = A H^+ = P \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} Q^*.
\]

In recent years, motivated by applications in linear optics and in Procrustes problems, the concept of the polar decomposition has been generalized to spaces equipped with an indefinite scalar product. If \( A \in F^{n \times n} \), then the generalized polar decomposition with respect to a given scalar product on \( F^n \) is defined to be a decomposition of the form \( A = W S \), where \( W \) is an automorphism with respect to the scalar product and \( S \) is a selfadjoint matrix whose nonzero eigenvalues are contained in the open right half-plane.

In contrast to the polar decomposition, the generalized polar decomposition of a matrix \( A \) need not exist, and if it exists, then the selfadjoint factor \( S \) is unique if \( A \) is nonsingular, but need not be unique if \( A \) is singular. Moreover, so far the concept of polar decomposition has been generalized to the case of square matrices only. Therefore, the question arises how the concept can be generalized to rectangular matrices and how uniqueness of both factors can be enforced.

In this talk, we introduce the canonical generalized polar decomposition \( A = W S \) defined for a general \( m \times n \) matrix \( A \) with respect to two scalar products on \( F^m \) and \( F^n \), respectively, defined by two nonsingular matrices \( M \in F^m \) and \( N \in F^n \). Our definition is based on the concept of partial \((M, N)\)-isometries which itself requires an appropriate generalization of the Moore-Penrose pseudoinverse. We derive conditions under which a unique decomposition exists and show how these decompositions can be computed by matrix iterations.
Smooth SVD Methods for the Computation of Sacker-Sell Spectra

Vu Hoang Linh and Volker Mehrmann

Abstract

This talk is devoted to the numerical approximation of Lyapunov and Sacker-Sell spectral intervals for linear differential-algebraic equations (DAEs). The spectral analysis for DAEs is improved and the concepts of leading directions and solution subspaces associated with spectral intervals are extended to DAEs. Numerical methods based on smooth singular value decompositions are introduced for computing all or only some spectral intervals and their associated leading directions. The numerical algorithms as well as implementation issues are discussed in detail and numerical examples are presented to illustrate the theoretical results.
Nearest Pencils with Specified Eigenvalues

Daniel Kressner, Emre Mengi, Ivica Nakic and Ninoslav Truhar

Abstract

We consider the distance from a linear pencil $A - \lambda B$ to a nearest pencil with specified eigenvalues. Specifically let $A, B \in \mathbb{C}^{n \times m}$ with $n \geq m$, $\lambda_1, \ldots, \lambda_k \in \mathbb{C}$ be given complex scalars and $r$ be a given positive integer. Denote the algebraic multiplicity of $\lambda_j$ as an eigenvalue of the pencil $A - \lambda B$ by $m_j(A,B)$ (possibly zero if $\lambda_j$ is not an eigenvalue) for $j = 1, \ldots, k$. We focus on the quantity

$$
\tau_r(A,B) = \inf \left\{ \left\| \begin{bmatrix} \delta A & \delta B \end{bmatrix} \right\|_2 : \sum_{j=1}^k m_j(A + \alpha_A \delta A, B + \alpha_B \delta B) \geq r \right\}.
$$

We derive a singular value optimization characterization when $\alpha_A = 1$ and $\alpha_B = 0$. We conjecture a singular value characterization when $\alpha_A = \alpha_B = 1$.

The starting point for our derivation is the Sylvester equation

$$AX - BXC(\mu, \Gamma) = 0,$$

where $C(\mu, \Gamma) = \begin{bmatrix} \mu_1 & -\gamma_1 \gamma_{12} & \cdots & -\gamma_1(r-1) & -\gamma_1r \\ 0 & \mu_2 & \cdots & -\gamma_2(r-1) & -\gamma_2r \\ \vdots & & & & \\ 0 & 0 & \cdots & \mu_{r-1} & -\gamma_{(r-1)r} \\ 0 & 0 & \cdots & 0 & \mu_r \end{bmatrix} \in \mathbb{C}^{r \times r},$

and $\mu, \Gamma$ are vectors consisting of $\mu_j, \gamma_{j\ell}$, respectively. For generic values of $\Gamma$ we have $\sum_{j=1}^k m_j(A,B) \geq r$ if and only if the solution space of the Sylvester equation is of dimension at least $r$ for some $\mu_1, \ldots, \mu_r \in \{\lambda_1, \ldots, \lambda_k\}$. Consequently, we obtain a singular value characterization, when $\alpha_A = 1$ and $\alpha_B = 0$, of the form

$$
\tau_r(A,B) = \inf \sup_{\mu} \sigma_{mr-r+1} \left( I \otimes A - C(\mu, \Gamma)^T \otimes B \right)
$$

by the Kroneckerization of the Sylvester equation.

Our initial interest in this problem was due to an inverse image processing application suggested by Golub and his colleagues [1]. In computer aided tomography, geophysics and other applications, several moments over a region can be measured. The problem is then to estimate the shape of the region. In particular if the shape is assumed to be a polygon, the vertices of the polygon turn out to be the eigenvalues of a rectangular pencil $A - \lambda B$ where $A, B$ are Hankel matrices defined in terms of moments. Due to measurement errors in moments the rectangular pencil has no eigenvalue, so one approach is to determine nearby pencils with full sets of eigenvalues. This is a special case of the problem considered here.

Additionally, many classical unstructured matrix nearness problems are special cases of the problem considered here, e.g., distance to instability, distance to uncontrollability, distance to the nearest matrix with a multiple eigenvalue.

References

Linear and Nonlinear Eigenvalue Problems in the Analysis and Robust Control of Time-delay Systems

Wim Michiels

Abstract

Time-delays are important components of many systems from engineering, economics and the life sciences, due to the fact that the transfer of material, energy and information is mostly not instantaneous. They appear for instance as computation and communication lags, they model transport phenomena and heredity and they arise as feedback delays in control loops. From a qualitative point of view the presence of time-delays in dynamical systems may induce complex behavior, and this behavior is not always intuitive. Time-delays in control loops are usually associated with degradation of performance and robustness, but there are situation where time-delays are beneficial and even used as controller parameters. They may also interact with different scales of the system: sometimes very large delays can be tolerated, but there are situations where an arbitrarily small delay may destabilize a stable system. In my presentation I will give an overview of the recent work in my group on eigenvalue based methods for stability analysis and robust control of linear time-delay systems.

The solutions of the linear delay differential equation

\[ \dot{x}(t) = A_0 x(t) + \sum_{i=1}^{m} A_i x(t - \tau_i) \]  

satisfy a spectrum determined growth property, in the sense that the asymptotic growth rate of solutions and the stability properties of the null solution are determined by the solutions of the nonlinear eigenvalue problem

\[ (\lambda I - A_0 - \sum_{i=1}^{m} A_i e^{-\lambda \tau_i}) v = 0, \quad v \neq 0. \]  

In the first part of my talk I will outline methods to compute the rightmost eigenvalues. Instrumental to this is the property that (2) can be equivalently rewritten as the standard eigenvalue problem

\[ (\lambda I - A) \varphi = 0, \quad \varphi \neq 0, \]  

where \( A \) is a linear operator acting on an infinite-dimensional function space to which \( \varphi \) belongs. This naturally leads to a combined approach: discretizing the linear infinite-dimensional eigenvalue problem (3) allows to obtain global information (e.g., estimates of all eigenvalue close to a target or in a prescribed region), while the finite-dimensional nonlinear problem (2) can be used for local corrections. Both direct and iterative methods will be discussed. It will also be outlined how these methods can be extended to a broad class of nonlinear eigenvalue problems.

In the second part of my presentation I will discuss the concept and computation of pseudospectra and stability radii for time-delay systems. Also here the two viewpoints (2)-(3) play a role: the transient growth of solutions is related to the pseudospectra of the linear operator \( A \). More important in the context of robust control are the structured pseudospectra assessing the effect of concrete perturbation on the system matrices in (1). It will be shown that they can be computed from (2) in a similar way as in Tisseur and Higham’s work on polynomial eigenvalue problems [7].
Finally, predictor-correct methods for the computation of $\mathcal{H}_\infty$ norms and pseudospectral abscissa will be briefly reviewed.

In the last part of my presentation I will present new control design methods for problems where the controller structure or order is a-priori specified (e.g., imposed from practical considerations) and the closed-loop system (plant + controller) can be expressed in the form

$$E\dot{z}(t) = A_0(p)z(t) + \sum_{i=1}^{m} A_i(p)z(t-\tau_i),$$

where $p$ is a vector corresponding to the controller parameters. These methods are based on a direct optimization of appropriately defined cost functions as a function of the parameters $p$. They are inspired by recent work on low-order control design for finite-dimensional systems (see, e.g., [1]) and rely on solving eigenvalue optimization problems. The analysis and design problems under consideration include the stabilization problem and the computation and optimization of $\mathcal{H}_2$ and $\mathcal{H}_\infty$ type cost functions. These new methods bridge the gap between the two mainstream approaches for control of linear time-delay systems, which are based on applying the systems theory for infinite-dimensional systems to a reformulation of the system equations in a first-order form, and based on a finite-dimensional approximation of the system, respectively.

References


PDE eigenvalue problems of a general form

\[ \mathcal{L}(\lambda, u) = 0 \]

with partial differential operator \( \mathcal{L} \), arise in many modern technological applications, e.g. vibration of structures or quantum phase transitions [7, 9]. Recently, a particular interest has been concerned with respect to so-called \textit{Adaptive Finite Element Methods} (AFEM) which, based on the quality of the numerical approximation (a posteriori error estimator), automatically adjust the finite element space in order to determine the sufficiently accurate final solution.

A typical loop of the AFEM consists of the four steps

**Solve \( \rightarrow \) Estimate \( \rightarrow \) Mark \( \rightarrow \) Refine.**

A large number of results on a posteriori error analysis, convergence and quasi-optimality of AFEM algorithms dedicated for eigenvalue problems were proposed during the last years, e.g. [4]. Except of very few works, e.g., [1, 6] for boundary value problems or [3] for eigenvalue problems, the resulting finite dimensional algebraic problem (linear system or eigenvalue problem) is assumed to be solved exactly and the computational costs as well as the fact that this problem is solved in finite precision arithmetic are typically ignored. Although determining a Galerkin solution for boundary value problems, on currently available computers, is not regarded as a problem, the situation is completely different if we think about eigenvalue problems.

We will first introduce an extended approach AFEMLA [8] for the adaptive finite element solution of selfadjoint elliptic PDE eigenvalue problems that incorporates the solution (in finite precision arithmetic) of the algebraic problem into the adaptation process and uses an early terminated iterative Krylov subspace method to compute a few smallest eigenvalues of a selfadjoint elliptic PDE eigenvalue problems. Using perturbation results for the generalized eigenvalue problems, i.e., the backward error analysis and the saturation assumption, we will prove our observations theoretically obtaining the bounds for the eigenvalues and eigenfunctions (eigenvectors).

An additional advantage of the proposed algorithm over the standard AFEM is that this adaptivity technique can be applied even without explicit knowledge of the underlying PDE. Let us assume that we would like to obtain the solution of some physical problem, e.g., compute the noise level inside the car, but the only available information are the matrix representation of the problem and the corresponding finite element grid. Although we do not have a PDE equation describing the problem and we are not able to construct appropriate an a posteriori error estimator, we still can construct an adaptive algorithm which allows us to obtain a good approximation of the exact solution at a reasonable cost. Since in the AFEMLA algorithm adaptivity is governed by the algebraic residual it can be used also when the problem come in discretized form, e.g. from the finite element modeling.

In [2] Arioli et. all. introduced functional backward errors and so-called compatibility theorem for boundary value problems. We describe the extension of this result to the eigenvalue problems.
Moreover, with help of a standard residual-type a posteriori error estimator $\eta(\lambda_h, u_h)$ and a discrete equivalence of the norm of the residual in the dual space $H^{-1}(\Omega)$ proposed in [5], we derive a combined a posteriori error estimator for a global error in the $H^1(\Omega)$-norm, i.e.,

$$
\|u - \bar{u}_h\|_{H^1(\Omega)} \leq \|u - u_h\|_{H^1(\Omega)} + \|u_h - \bar{u}_h\|_{H^1(\Omega)} \lesssim \eta(\lambda_h, u_h) + \|r\|_{H^{-1}},
$$

where $\|u - u_h\|_{H^1(\Omega)}$ is the discretization error and $\|u_h - \bar{u}_h\|_{H^1(\Omega)}$ the algebraic error. Several numerical examples will be presented to illustrate the performance of the algorithm. A proper equilibration between the discretization and iteration error will be discussed to derive an efficient stopping criteria for the iterative eigensolver.

Since, a long term goal of our research is to design adaptive methods for non-selfadjoint, parameter dependent problems, we will consider some possible extensions of presented results.

References


Accurate Solution of Structured Linear Systems and Least Square Problems Through Rank Revealing Decompositions

Johan A. Ceballos, Froilán M. Dopico, Juan M. Molera

Abstract

Linear systems of equations $Ax = b$ and Least Square Problems $Ax \simeq b$ where the matrix $A$ has some particular structure arise frequently in applications. Very often structured matrices have huge condition numbers $\kappa(A)$ and, therefore, standard algorithms fail to compute accurate solutions of $Ax = b$ or $Ax \simeq b$. In this work, we introduce a framework that allows us to solve accurately many classes of structured systems, both linear and least squares, independently of the condition number of $A$ and in an efficient way, with cost $O(n^3)$. For most of these classes, no algorithms are known that are always accurate and efficient. The approach in this work relies on computing first an accurate rank-revealing decomposition of $A$, an idea that has been widely used in the last decades to compute singular value and eigenvalue decompositions of structured matrices with high relative accuracy. The proposed method can be applied to many well known classes of structured matrices. Among others, it can be applied to Cauchy and Vandermonde matrices, with any distribution of nodes, i.e., without requiring $A$ to be totally positive, and Graded matrices, diagonal scalings of a well-conditioned matrix.
A New Approach to Nonsymmetric Lanczos and to Avoiding Breakdown

Ron Morgan

Abstract

The nonsymmetric Lanczos algorithm is often used for solving large systems of linear equations. It is the basis for such popular algorithms as BiCGStab, TFQMR, and more recently IDR. However for the task of computing eigenvectors, nonsymmetric Lanczos is not popular. This is both because of the large amount of storage that is needed for a non-restarted Krylov iteration and because of roundoff error in the nonsymmetric Lanczos iteration. Instead, restarted Arnoldi methods such as implicitly restarted Arnoldi [6] are the preferred Krylov methods. Here we discuss a new approach to nonsymmetric Lanczos that deals with the storage problem by restarting and also addresses the roundoff error problems (including alternatives to look-ahead for near-breakdown). This new method is an alternative to Arnoldi for computing right and left eigenvectors. It is particularly useful for sparse matrices that are not highly non-normal.

Nonsymmetric Lanczos with deflated restarting

For previous work on restarted Arnoldi and deflated GMRES, see for example [5, 6, 2, 3, 1]. The new method is called nonsymmetric Lanczos with deflated restarting or NLan-DR. It uses the nonsymmetric Lanczos recurrence, but has restarting to limit the dimensions of the subspaces. At the restart, right and left Ritz vectors are computed and used as starts for the subspaces in the next cycle. The new subspaces are generated with the Lanczos recurrence, modified for the first couple iterations. The new right subspace is

\[ \text{Span}\{y_1, y_2, \ldots, y_k, Ay_1, A^2y_1, A^3y_1, \ldots, A^{m-k}y_1, \} \],

for each right Ritz vector \( y_i \). So this subspace contains Krylov subspaces with each Ritz vector as starting vector. The left subspace is similar but with left Ritz vectors. NLan-DR needs some reorthogonalization, but usually less than Arnoldi.

Example 1. We use a bidiagonal test matrix of size \( n = 2500 \). The diagonal entries are 0.1, 0.2, 0.3, 0.4, 1, 2, 3, \ldots, 2495, 2496 and the superdiagonal has all 0.1’s. NLan-DR saves 15 right and left Ritz vectors at the restart and allows subspaces to grow out to size 60. Periodic rebiorthogonalization of two consecutive vectors is done every 15 iterations. The convergence is almost identical to that of restarted Arnoldi and the cost is considerably cheaper. NLan-DR requires 5620 vector operations per cycle, while two runs of restarted Arnoldi (for right and left eigenvectors) total 17460 per cycle.

Near-breakdown control

Near-breakdown in the Lanczos iteration can occur when corresponding right and left Lanczos vectors are nearly orthogonal. This causes significant roundoff error. Look-ahead Lanczos [4] is the well-known treatment, but it is not easy to implement. The fact that NLan-DR has restarting offers the opportunity to avoid near-breakdown in new ways. We first take an approach of monitoring the angle between corresponding right and left Lanczos vectors and restarting when the angle is too small. However, we go back a couple of iterations before restarting to avoid having the same small angle appear.
Example 2. The matrix is the same as for the first example except off-diagonal elements are all 1’s. For some choices of starting vector, NLan-DR performs well, but for others, near-breakdown causes large error. For a particularly bad case, some residual norms only reach relative accuracy of $10^{-5}$. This is because the cosine of the angle between the right and left eigenvectors hits $3.0 \times 10^{-6}$ at one point. With the approach mentioned above of going back two iterations and restarting, residual norms relative to the matrix norm reach almost $10^{-10}$.

This approach to near-breakdown control is not effective for highly non-normal matrices. We now mention some other approaches that we have begun investigating. At the restart, a switch can be made to harmonic Rayleigh-Ritz. A generalization of this would be to adapt Stewart’s Krylov-Schur [7] to nonsymmetric Lanczos. Another option is at the point of near-breakdown, to restart and change to a different inner-product that is tailored to push the Lanczos vectors away from near orthogonality. We are also interested in the question of why near-breakdown seems to occur at the same point even when different inner-products are used from the onset of Lanczos.

**Linear equations with multiple right-hand sides**

An application that needs both right and left eigenvectors is the solution of systems of linear equations with multiple right-hand sides. The first system can be solved with NLan-DR (simultaneously with the computation of right and left eigenvectors). Then subsequent right-hand side systems can be solved with a deflated Bi-CGStab that has a projection using the right and left eigenvectors, then application of Bi-CGStab. For the matrix of Example 2, the iterations for a subsequent right-hand side can be reduced by almost an order of magnitude.

**References**


Directional Perturbation in Structured Eigenproblems

Julio Moro
Departamento de Matemáticas
Universidad Carlos III de Madrid

Abstract

The design and analysis of structure-preserving algorithms to solve structured eigenproblems has led in the last decades to a steady interest in structured spectral perturbation theory, i.e. in determining the behavior of eigenvalues and other spectral objects (e.g., invariant subspaces, sign characteristics,...) when a matrix or, more generally, an operator is subject to perturbations which belong to the same class of operators as the unperturbed one. It is well known that this behavior is usually quite different from the behavior under arbitrary, nonstructured perturbations. In this talk we give an overview of first order structured eigenvalue perturbation theory, i.e. of results involving the local variation of eigenvalues as expressed by their directional derivatives, and their interaction with the fact that perturbations are restricted within the class of interest. Such results are relevant in many practical situations when eigenvalues need to be pushed in certain specific directions, or must be moved as fast as possible away from a critical (or dangerous) region by a small, usually structured, perturbation. Special emphasis is made on classes of matrices and matrix pencils with symmetries in some indefinite scalar product, which often arise in several applications in Control and Systems Theory.
A Computational Approach for Large Scale Nonlinear Least Squares Problems

Qing Chu, Ying-Wai Fan, and James Nagy

Abstract

In this talk we consider an approach to solve the large scale nonlinear least squares problem

$$\min_{w_i, f} \left\| \begin{bmatrix} g_1 \\ \vdots \\ g_m \end{bmatrix} - \begin{bmatrix} K(w_1) \\ \vdots \\ K(w_m) \end{bmatrix} f \right\|_2^2 = \min_{w_i, f} \left\{ \|g_1 - K(w_1)f\|_2^2 + \cdots + \|g_m - K(w_m)f\|_2^2 \right\}$$

(1)

where $g_i$, $w_i$, and $f$ are vectors that can have $n = 10^4$ or more elements, and $K(w_i)$ are $n \times n$ sparse and/or structured matrices. A standard approach to solve nonlinear least squares problems of the form (1) is the variable projection method [3, 5, 6, 7, 8, 9, 10], where the linear term $f$ is mathematically eliminated from the minimization, and a Gauss-Newton method is used to compute the nonlinear terms $w_i$. A computational challenge is that the coupling of $w_i$ results in a large dense Jacobian of the reduced cost functional. To get around this difficulty we reformulate the minimization problem (1) through a decoupling scheme [4]. In particular, we solve the minimization problem

$$\min_{w_i, f_i} \left\{ \|g_1 - K(w_1)f_1\|_2^2 + \|g_2 - K(w_2)f_2\|_2^2 + \cdots + \|g_m - K(w_m)f_m\|_2^2 \\
+ \|f_1 - f_2\|_2^2 + \|f_2 - f_3\|_2^2 + \cdots + \|f_{m-1} - f_m\|_2^2 + \|f_m - f_1\|_2^2 \right\},$$

With this decoupling the Jacobian matrix has a sparse, block structure, and it is possible to solve the Jacobian system using an iterative least squares method, such as LSQR. Furthermore, convergence of the Gauss-Newton method for the decoupled problem is only slightly slower than for the coupled problem.

We illustrate the performance of this scheme using a specific example in adaptive optics imaging where $w_i \equiv w_i(x,y)$ are wavefront functions of light incoming to the telescope mirror during the detection of image $g_i$. We show how good initial guesses of $w_i$ can be obtained by solving two linear least squares problems [1, 2], and describe how to efficiently implement linear solvers involving the Jacobian system [4].

References

http://www.mathcs.emory.edu/technical-reports/techrep-00205.pdf


http://www.mathcs.emory.edu/technical-reports/techrep-00204.pdf


Optimizing Halley’s Iteration for Computing the Matrix Polar Decomposition

Yuji Nakatsukasa, Zhaojun Bai and François Gygi

Abstract
Computing the matrix polar decomposition $A = UH$, where $U$ is unitary and $H$ is a Hermitian positive semidefinite, is required in many applications. Our motivation is from solving a large-scale orthogonal Procrustes problem arising from the subspace alignment in the first-principles molecular dynamics simulations of electronic structure calculations [1, 3].

The most well-known algorithm for computing the polar decomposition is the scaled Newton iteration

$$X_{k+1} = \frac{1}{2} \left( \zeta_k X_k + (\zeta_k X_k)^{-H} \right), \quad X_0 = A, \quad (1)$$

where $\zeta_k$ is a scaling factor [4]. With the suboptimal scaling presented in [2] the iteration (1) converges to the unitary polar factor $U$ with the tolerance $10^{-16}$ within 9 iterations if the condition number $\kappa_2(A) \leq 10^{16}$. It is proven that the method is backward stable if the matrix inverses are computed in a forward-backward stable manner [5, 2].

Computing the matrix inversion is expensive in communication costs. On the emerging multicore and heterogeneous computing systems, communication costs have exceeded arithmetic costs by orders of magnitude, and the gap is growing exponentially over time. This motivated our study on inverse-free methods for computing the matrix polar decomposition. Previous studies exist in this direction, but have encountered either backward instability or slow convergence compared with the scaled Newton iteration.

In this talk, we first present our recent work on a new inverse-free polar decomposition algorithm [6]. The new method can be mathematically expressed as a generalization of Halley’s iteration:

$$X_{k+1} = X_k (a_k I + b_k X_k^H X_k) (I + c_k X_k^H X_k)^{-1}, \quad X_0 = A/\|A\|, \quad (2)$$

where the scaling parameters $a_k$, $b_k$ and $c_k$ are dynamically chosen in a suboptimal way so as to speed up the convergence and meanwhile improve the backward stability. We show that the iteration (2) needs at most 6 iterations for convergence to $U$ with the tolerance $10^{-16}$ for any matrix $A$ with $\kappa_2(A) \leq 10^{16}$.

For practical implementation, the iteration (2) can be written in the following mathematically equivalent form:

$$\begin{align*}
X_{k+1} &= \frac{\sqrt{c_k} X_k}{I} = \left[ \begin{array}{c} Q_1 \\ Q_2 \end{array} \right] R \quad \text{(QR decomposition)}, \\
X_{k+1} &= \frac{b_k}{c_k} X_k + \frac{1}{\sqrt{c_k}} \left( a_k - \frac{b_k}{c_k} \right) Q_1 Q_2^H, \quad X_0 = A/\|A\|. \quad (3)
\end{align*}$$

The QR decomposition-based nature of (3) makes it a communication-friendly method and offers an attractive alternative to the scaled Newton iteration on multicore and heterogeneous parallel computing architectures.

In the second half of this talk, we plan to present our progress on the implementation and performance benchmarking of the inverse free method, and a theoretical proof of the backward stability, which shows that the choice of new scaling parameters plays an important role in the backward stability.
References


A Combinatorial Problem in Sparse Orthogonal Factorization

Esmond G. Ng
Lawrence Berkeley National Laboratory

Abstract

Let \( A \) be a large sparse \( m \times n \) matrix and \( b \) be an \( m \)-vector, with \( m \geq n \). We consider the solution of the least squares problem

\[
\min_x ||Ax - b||_2
\]

(1)

using orthogonal factorization. Denote the orthogonal factorization of \( A \) by

\[
A = Q \begin{bmatrix} R \\ O \end{bmatrix},
\]

where \( Q \) is an \( m \times m \) orthogonal matrix and \( R \) is an \( n \times n \) upper triangular matrix. We assume that \( A \) has full rank, so that \( R \) is nonsingular. It is well known that there will be fill during the factorization; that is, some of the zero entries in \( A \) will become nonzero. Here, we are particularly interested in the amount of fill in the upper triangular factor \( R \). The amount of fill in \( R \) depends on the ordering of the columns of \( A \). We will therefore assume that the columns of \( A \) have been appropriately reordered using a fill-reducing ordering algorithm. If we are able to compute \( Q \) and \( R \), then the least squares problem in (1) can be solved easily. However, because of fill, it may be expensive to compute \( Q \) and \( R \).

Suppose we partition the matrix \( A \) and the vector \( b \) into

\[
A = \begin{bmatrix} A_s \\ A_d \end{bmatrix}, \quad b = \begin{bmatrix} b_s \\ b_d \end{bmatrix},
\]

where \( A_s \) is \( m_s \times n \), \( A_d \) is \( m_d \times n \), and \( b_s \) and \( b_d \) are of lengths \( m_s \) and \( m_d \), respectively, with \( m_s + m_d = m \). Instead of computing the complete orthogonal factorization of \( A \), let’s assume that we only compute the orthogonal factorization of \( A_s \). Denote the orthogonal factorization of \( A_s \) by

\[
A_s = Q_s \begin{bmatrix} R_s \\ O \end{bmatrix},
\]

where \( Q_s \) is an \( m_s \times m_s \) orthogonal matrix and \( R_s \) is an \( n \times n \) upper triangular matrix. We will again assume that \( A_s \) has full rank so that \( R_s \) is nonsingular. As we will see below, a desirable property is for \( R_s \) to have fewer nonzero entries and require fewer operations to compute than \( R \).

Suppose that \( y_{\min} \) is the vector that solves the least squares problem

\[
\min_y ||A_s y - b_s||_2.
\]

It has been shown in [3] that \( R_s, A_d, \) and \( b_d \) can be used to update \( y_{\min} \) to obtain a solution to (1). We refer to this as the direct approach. See [1,3] when \( A_s \) is rank deficient.

We have also shown recently in [2] that \( R_s \) can be used to precondition the least squares problem in (1):

\[
\min_x ||AR_s(R_s^{-1}x) - b||_2.
\]

(2)
In particular, if (2) is solved using LSQR [4], then the number of iterations required is bounded by $m_d$, the number of rows in $A_d$, at least in theory. We refer to this as the iterative approach. If $A_s$ is rank deficient, see [2] for further details.

In both the direct and iterative approaches, we are using the orthogonal factorization of $A_s$, which is a submatrix of $A$. It is desirable for $R_s$ to be sparser than $R$ so that the cost of applying $R_s$ (in triangular solutions) can be kept small. Also, it is important that $R_s$ can be computed much more efficiently than $R$, both in terms of storage and time, so that the cost of its computation does not dominate the cost of the entire solution process. This gives rise to the following combinatorial problem – determine a submatrix $A_s$ from $A$ so that the orthogonal factorization of $A_s$ requires a small number of operations and produces a sparse upper triangular factor $R_s$. In our context, we would like to identify rows of $A$ that would introduce significant amount of fill in $R_s$ if they were included in $A_s$. Such rows are often referred to as “dense” rows and will be included in $A_d$. In this talk, we consider a number of heuristics for tackling this combinatorial problem.

This is joint work with Barry Peyton at Dalton College.

References


Euclidean Distance Matrix Completion Problems and Protein Structure Determination

Haw-ren Fang and Dianne P. O’Leary

Abstract

Given the distances between each pair of \( n \) particles in \( R^r, n \geq r \), it is easy to determine the relative positions of the particles. In many applications, though, we are given only some of the distances and we would like to determine the missing distances and thus the particle positions. For example, measurement by nuclear magnetic resonance (NMR) spectroscopy on a protein molecule can be used to obtain accurate estimates of distances between atoms that are close, but gives little information about large distances.

We focus in this work on algorithms to solve this distance completion problem.

A Euclidean distance matrix is one in which the \((i,j)\) entry specifies the squared distance between particle \( i \) and particle \( j \). Given a partially-specified symmetric matrix with zero diagonal, the Euclidean distance matrix completion problem (EDMCP) is to determine the unspecified entries to make the matrix a Euclidean distance matrix.

We survey three different approaches to solving the EDMCP. We advocate expressing the EDMCP as a nonconvex optimization problem using the particle positions as variables and solving using a modified Newton or quasi-Newton method. To avoid local minima, we develop two approaches: a randomized initialization technique that involves a nonlinear version of the classical multidimensional scaling, and a dimensionality relaxation scheme with optional weighting.

Our experiments show that the method easily solves the artificial problems introduced by Moré and Wu. It also solves the 12 much more difficult protein fragment problems introduced by Hendrickson, based on a typical small protein, bovine pancreatic ribonuclease A, which consists of 124 amino acids, and after discarding end chains, 1849 atoms. We also solve the 6 larger protein problems introduced by Grooms, Lewis, and Trosset.
Representation of high-dimensional arrays (tensors) plays increasingly important role in applications, where solution of high-dimensional problems is required. Due to the curse of dimensionality, one has to approximate such arrays with some low-parametric representation. Recently, Tensor Train (TT) decomposition was introduced as an alternative to canonical and Tucker formats for the approximation of high-dimensional arrays (tensors). It is shown that it avoids the curse of dimensionality in the case when good low-tensor rank approximation exists, but in several cases it can have even much smaller number of parameters. In can be computed via stable numerical procedure (TT-SVD), thus it is good generalization of singular value decomposition to high-dimensional arrays. In this talk we will present basic concept for work with the TT-format: transformation of the tensor into the TT-format, basic linear algebra, matrix-by-vector product, tensor rounding (approximation of a given TT-tensor with another one with smaller number of parameters, but within prescribed accuracy). This machinery allows us to develop numerical methods for the solution of important high-dimensional problems from quantum chemisty (quantum molecular dynamics) and stochastic PDE using the same tools. Below we give a brief “TT cheatsheet” which describes the format itself, and how basic operations are performed. They are very simple and quite elegant, and often take several tens of lines of MATLAB code to program. The main object is a $d$-dimensional array $A$. For tensor $A$ its elements are denoted as $A(i_1,\ldots,i_d)$.

**TT-format**

$A(i_1,\ldots,i_d) = G_1(i_1)\ldots G_d(i_d)$, $G_k(i_k)$ is $r_{k-1}\times r_k$ matrix, $r_0 = r_d = 1$.

**Addition**

$C = A + B :$

$C_k(i_k) = \begin{pmatrix} A_k(i_k) & 0 \\ 0 & B_k(i_k) \end{pmatrix}$, $k = 2,\ldots,d - 1$,

$C_1(i_1) = \begin{pmatrix} A_1(i_1) \\ B_1(i_1) \end{pmatrix}$, $C_d(i_d) = \begin{pmatrix} A_d(i_d) \\ B_d(i_d) \end{pmatrix}$.

**Hadamard**

$C = A \odot B :$

$C_k(i_k) = A_k(i_k) \odot B_k(i_k)$.

**Matrix in TT**

$M(i_1,\ldots,i_d; j_1,\ldots,j_d) = M_1(i_1,j_1)\ldots M_d(i_d,j_d)$, $M_k(i_k,j_k)$ is $r_{k-1}\times r_k$ matrix, $r_0 = r_d = 1$.

**Matvec**

$Y = MX$, $M$ — matrix with cores $M_k(i_k,j_k)$, $X$ has cores $X_k(j_k)$

$Y_k(i_k) = \sum_{j_k} M_k(i_k,j_k) \odot X_k(j_k)$.

**Contraction**

$W = A \times_1 u_1 \ldots \times_d u_d$,

$\Gamma_k = \sum_{i_k} A_k(i_k) u_k(i_k)$, $W = \Gamma_1\ldots\Gamma_d$. 

175
Fast Algorithms for the Approximation of the Pseudospectral Abscissa and Pseudospectral Radius of a Matrix

Nicola Guglielmi and Michael L. Overton

Abstract

The \( \epsilon \)-pseudospectral abscissa and radius of an \( n \times n \) matrix are respectively the maximal real part and the maximal modulus of points in its \( \epsilon \)-pseudospectrum, defined using the spectral norm. Existing techniques compute these quantities accurately but the cost is multiple singular value decompositions and eigenvalue decompositions of order \( n \), making them impractical when \( n \) is large. We present new algorithms based on computing only the spectral abscissa or radius of a sequence of matrices, generating a sequence of lower bounds for the pseudospectral abscissa or radius. We characterize fixed points of the iterations, and we discuss conditions under which the sequence of lower bounds converges to local maximizers of the real part or modulus over the pseudospectrum, proving a locally linear rate of convergence for \( \epsilon \) sufficiently small. The convergence results depend on a perturbation theorem for the normalized eigenprojection of a matrix as well as a characterization of the group inverse (reduced resolvent) of a singular matrix defined by a rank-one perturbation. The total cost of the algorithms is typically only a constant times the cost of computing the spectral abscissa or radius, where the value of this constant usually increases with \( \epsilon \), and may be less than 10 in many practical cases of interest.
A Backward Rounding Error Analysis of the Lanczos Process, and its Implications for Other Krylov Subspace Methods

Chris Paige

Abstract

Here we show that at step $k$ a well implemented Lanczos process for tridiagonalizing an $n \times n$ Hermitian matrix $A$, with initial vector $v_1$, gives a real symmetric tridiagonal matrix $T_k$ that is the exact result of applying the Lanczos process to the $(k+n) \times (k+n)$ Hermitian matrix $\text{diag}(T_k, A) + H_k$, with initial vector $(0, v_1^T)^T$, where $\|H_k\|_2 \leq O(\epsilon)\|A\|_2$ and $\epsilon$ is the computer floating point precision. See [1, Theorem 3.1]. Thus the computed tridiagonal matrix could be thought of as the “backward stable” result of a strange, larger-dimensioned problem.

Could this property be called “Augmented backward stability”?

The supposedly orthonormal vectors $\tilde{v}_j$ produced by the finite precision process can quickly lose orthogonality, but are all nearly unit vectors (i.e. of 2-norm one). When each $\tilde{v}_j$ is normalized to the corresponding $v_j$ having unit norm, giving $V_k = [v_1, \ldots, v_k]$, these vectors appear in the orthonormal vectors in the above larger dimensioned exact Lanczos process. If $V_k^H V_k = I + U_k^H + U_k$, with $U_k$ strictly upper triangular, and $S_k \equiv (I + U_k)^{-1} U_k$, then $\|S_k\|_2 \leq 1$, and the number of unit singular values of $S_k$ is exactly the rank deficiency of $V_k$, and so of $V_k \equiv [\tilde{v}_1, \ldots, \tilde{v}_k]$, [1, Corollary 2.2]. Thus $S_k$ contains all we need to know about the loss of orthogonality and rank of $V_k$. The exact $(k+n) \times (k+n)$ Lanczos recurrence for $T_k$ also contains $S_k$ as part of the vector recurrence, and so we not only have the true development of $T_k$, but of the loss of orthogonality matrix $S_k$ too.

This analysis is a step towards a deeper understanding of the Lanczos process, and we illustrate some of the insights so far gained. It is intended to be a basis for developing analyses of how the Lanczos process behaves when combined with other computations for finding eigenvalues and singular values and solving linear systems of equations and linear least squares problems in methods such as conjugate gradients (CG), SYMMLQ, MINRES, LSQR, etc., and in related algorithms such as the unsymmetric Lanczos process.

It might also form a template for analyzing the behaviors of other Krylov subspace methods which are obviously not backward stable in the usual sense, but which are nevertheless very useful for large sparse matrix problems. Suppose we can obtain such an “augmented backward stability” result for a good Krylov subspace method. The advantages of this would in some way parallel Wilkinson’s beautiful work on backward errors that revolutionized our understanding of “direct” methods, giving us the concept of backward stability, with its use in for example:

Size of backward error & Condition of problem $\rightarrow$ Bound on the solution error.  \hfill (1)

The “augmented backward stability” result for a good Krylov subspace method would show that the computed behavior at each step $k$ will be exact for $k$ steps of the method applied to a clearly stated problem of the correct form (such as the augmented Hermitian matrix for the Lanczos process mentioned above). As a result, the exact condition of, and convergence behavior of the method on, this new problem could be considered. This would be useful for understanding and bounding solution errors in a way similar to (1), but it might also be useful in the form:

Size of backward error & Form of “augmented” problem $\&$ Theoretical convergence properties $\rightarrow$ Actual convergence properties.

A key point is that the backward rounding error result would be derived independently of the problem condition, or convergence properties of the method, just as in [1], simplifying everything.
References

Ritz Vectors in the Lanczos Process

Chris Paige, Ivo Panayotov

Abstract

In this talk we will present a few remarkable properties of complex Hessenberg matrices based on some earlier ideas by J.-P. Zemke [3], and show how these can be used to analyze the length of Ritz vectors in the finite precision Lanczos process for computing eigenvalues of Hermitian matrices. This is a central question in the practical use of the Lanczos process since it is intimately related to the finite precision quality of eigenvalue estimation.

An analysis has already been performed on this topic, but only part of this has been published in a scientific journal, and that was a long time ago, see [1]. Our work provides a (hopefully) simpler analysis for the previously published case which applies to well separated Ritz values, as well as a new analysis for the remaining (unpublished) case which handles clusters of Ritz values. Our approach is more general than the approach in [1], in that it applies not only to the Lanczos process for real symmetric matrices, but to the Lanczos process adapted to any form of normal matrix with collinear eigenvalues—such as Hermitian and skew-Hermitian matrices. Much of the theory is even more general, so it may be useful in the analysis of other methods. For this reason we believe that the techniques and insights in our work are as important as the actual results. Thus, we will discuss a few key elements of these techniques in addition to presenting our results.

In our talk we assume that the audience is familiar with the Lanczos process for computing eigenvalues of Hermitian matrices, but do not assume any familiarity with the rounding error analysis of this process. We will present all the concepts needed to motivate our work and allow the audience to understand its relevance. The work has been submitted for publication in [2].

References


On Choice of Preconditioner for Minimum Residual Methods for Nonsymmetric Matrices

Jennifer Pestana, Andrew J. Wathen

Abstract

Large and sparse or structured linear or linearised systems of equations $Bx = b$, with real nonsymmetric coefficient matrix $B$, arise in a large range of applications. Such systems are frequently solved by iterative methods, of which GMRES [2] is one of the most popular. However, in many instances the standard convergence estimates for minimal residual (MR) methods, including GMRES, are found to be poor indicators of observed convergence. This is in contrast to iterative methods for real symmetric (or complex hermitian) matrices, for which convergence bounds—typically given in terms of polynomial approximation problems on the spectrum—are often fairly tight. As a consequence, it is difficult to determine what one is trying to achieve with preconditioning for nonsymmetric matrices: the glib answer ‘fast convergence’ gives no idea nor criteria for making other than heuristically motivated choices. We will present some limited theory which aims to provide some sufficient mathematical conditions for the selection of preconditioners in the nonsymmetric case.

We consider here the subset of real nonsymmetric coefficient matrices which are self-adjoint with respect to an inner product $\langle \cdot, \cdot \rangle_H$, or $H$-self-adjoint; these are precisely the matrices that are diagonalisable with real eigenvalues. When $B$ is $H$-self-adjoint it is possible to apply minimal residual methods in this $H$-inner product ($H$-MR methods) to $Bx = b$ (see section 13 in [3] for a survey). Moreover, the convergence of this $H$-MR method can be bounded by a quantity that depends only on the eigenvalues of $B$ [1]. That is, the $k$th residual $r_k$ satisfies

$$\frac{\|r_k\|_H}{\|r_0\|_H} \leq \min_{p \in \Pi_k, p(0) = 1} \max_{\lambda \in \sigma(B)} |p(\lambda)|,$$

(1)

where $\| \cdot \|_H$ is the norm induced by $\langle \cdot, \cdot \rangle_H$, $\Pi_k$ is the set of real polynomials of degree at most $k$ and $\sigma(B)$ is the spectrum of $B$.

We will show that convergence of this $H$-MR method can be related to that of standard MR methods, like GMRES, by including an extra factor that depends on the iteration number. In turn, this factor can be bounded by the spectral condition number of the matrix of eigenvectors, as in the standard eigenvalue-based MR bound. However, experiments suggest that when the coefficient matrix is $H$-self-adjoint, convergence of these standard MR methods (i.e., based on the Euclidean inner product) also depends essentially on the spectrum of the coefficient matrix, without the potentially huge contribution of the condition number of the eigenvector matrix.

This convergence bound is also applicable to left-preconditioned systems when the preconditioned coefficient matrix is self-adjoint with respect to an inner product. We show that if $B$ is $H$-self-adjoint and the preconditioner $P$ is self-adjoint and positive definite with respect to $\langle \cdot, \cdot \rangle_H$ then $P^{-1}B$ is self-adjoint with respect to the inner product $\langle \cdot, \cdot \rangle_{HP}$. In this way, self-adjointness can be preserved. It follows from (1) that convergence of $HP$-MR methods applied to this preconditioned system depend largely on eigenvalues. Specifically, the preconditioned residuals, $P^{-1}r_k^{HP}$, satisfy

$$\frac{\|P^{-1}r_k^{HP}\|_{HP}}{\|P^{-1}r_0^{HP}\|_{HP}} \leq \min_{p \in \Pi_k, p(0) = 1} \max_{\lambda \in \sigma(P^{-1}B)} |p(\lambda)|.$$
We present experiments that suggest that in this situation convergence of standard MR methods, such as GMRES, is also governed principally by eigenvalues.

We will furthermore show that when the restrictions on $P$ are relaxed slightly eigenvalues remain important in determining convergence. Specifically, when $P$ is self-adjoint and positive definite with respect to an inner product that is close to $\langle \cdot, \cdot \rangle_H$, say $\langle \cdot, \cdot \rangle_{H'}$, the following theorem provides a bound on the convergence of $HP$-MR methods applied to the preconditioned linear system.

**Theorem** Let $B$ be self-adjoint with respect to $\langle \cdot, \cdot \rangle_H$ and $P$ be self-adjoint and positive definite with respect to $\langle \cdot, \cdot \rangle_{H'}$. Let

$$P^{-1}B = (S - E)\Lambda(S - E)^{-1}$$

be a diagonalisation of $P^{-1}B$, where

$$HP = S^{-T}S^{-1}.$$ 

If $\|S^{-1}E\|_2 < 1$ the preconditioned residual $P^{-1}r^H_P$ of any minimum residual method in $\langle \cdot, \cdot \rangle_{HP}$ satisfies

$$\frac{\|P^{-1}r^H_P\|_{HP}}{\|P^{-1}r_0\|_{HP}} \leq \frac{1 + \|S^{-1}E\|_2}{1 - \|S^{-1}E\|_2} \min_{p \in \Pi_k, p(0) = 1} \max_{\lambda \in \sigma(P^{-1}B)} |p(\lambda)|.$$ 

It follows that, under the same assumptions in the theorem above, the convergence of $HP$-MR methods can be bounded by a quantity that depends only on eigenvalues to first order in $\|S^{-1}E\|_2$. That is,

$$\frac{\|P^{-1}r^H_P\|_{HP}}{\|P^{-1}r_0\|_{HP}} \leq (1 + O(\|S^{-1}E\|_2)) \min_{p \in \Pi_k, p(0) = 1} \max_{\lambda \in \sigma(P^{-1}B)} |p(\lambda)|.$$ 

We will present several examples which illustrate this theory and which indicate that the convergence of standard (i.e., Euclidean norm based) MR methods is similarly dependent on the eigenvalues of the preconditioned system. We discuss the implications of the above theorem for preconditioning a $H$-self-adjoint matrix and we comment on the case that $B$ is not self-adjoint with respect to an inner product but is self-adjoint with respect to a symmetric bilinear form. Since every matrix is self-adjoint with respect to a symmetric bilinear form [4], these comments apply to all real matrices.

**References**


The Total Least Squares Problem with Multiple Right-Hand Sides

Iveta Hnětynková, Martin Plešinger, Diana M. Sima, Zdeněk Strakoš, and Sabine Van Huffel

Abstract

The total least squares (TLS) techniques, also called orthogonal regression and errors-in-variables modeling, see [15, 16], have been developed independently in several areas. For a given linear (orthogonally invariant) approximation problem

$$AX \approx B,$$

where

$$A \in \mathbb{R}^{m \times n}, \quad B \in \mathbb{R}^{m \times d}, \quad X \in \mathbb{R}^{n \times d},$$

the TLS formulation aims at a solution of a modified problem

$$(A + E)X = B + G \quad \text{such that} \quad \min \| (G, E) \|_F.$$

The algebraic TLS formulation has been investigated for decades, see the early works [5], [4, Section 6], [13]. In [6] it is shown that even with $d = 1$ (which represents a problem with the single right-hand side $Ax \approx b$, where $b$ is an $m$-vector) the TLS problem may not have a solution and when the solution exists, it may not be unique. The classical book [14] introduces the generic-nongeneric terminology representing a commonly used classification of TLS problems. If $d = 1$, then the generic problems simply represent problems that have a (possibly nonunique) solution, whereas nongeneric problems do not have a solution. This is no longer true for $d > 1$. For $d > 1$, [14] analyzes only two representative cases characterized by the special distribution of singular values of the extended matrix $[B, A]$. A general case is not analyzed—it is considered only as a perturbation of one of the special cases. The so called classical TLS algorithm given in [14] computes some output $X$ for any data $A, B$, but the relationship of the output $X$ to the original problem is, however, not clear.

The single right-hand side problem has been recently revisited in a series of papers [9, 10, 11]. Here it is shown that the problem does not have a solution when the collinearities among columns of $A$ are stronger than the collinearities between $\mathcal{R}(A)$ and $b$. An analogous situation may occur for $d > 1$, but here different columns of $B$ may be correlated with different subsets of columns of $A$. Therefore it is no longer possible to stay with the generic-nongeneric classification of TLS problems. This is also the reason why the question remained open in [14]. In the first part of our contribution we try to fill this gap and investigate existence and uniqueness of the solution of the TLS problem with $d > 1$ in full generality. We suggest a classification of TLS problems revisiting and refining the basic generic-nongeneric terminology, see [12, 7].

A core reduction concept introduced in [11] makes a clear link between the original data and the output of the classical TLS algorithm for the problems with the single right-hand side (see also [7]). Therefore the core reduction is an appropriate tool for understanding the TLS problem with $d = 1$. In the second part of this contribution we introduce an extension of the core reduction for multiple right-hand sides problems. Following [11], we employ the SVD of $A$ which allows us to define the core problem and show its fundamental properties. Then we show how the core problem can be obtained by the band generalization of the Golub-Kahan iterative bidiagonalization algorithm (also called band-Lanczos algorithm) proposed for this purpose by Åke Björck in a series of lectures [1, 2, 3]. We show, together with other results in progress (see [12, 8]), that both approaches (based on the SVD of $A$ and on the band algorithm) give the same core problem up to an orthogonal transformation. Using the core reduction, we illustrate some particular difficulties which are present in the TLS problems with $d > 1$. 

182
References


A Perron Vector-based Iteration for Solving Quadratic Vector Equations

Dario A. Bini, Beatrice Meini, and Federico Poloni

Abstract

Markovian Binary Trees (MBTs) are a particular family of branching processes, which are used to model the growth of populations consisting of several types of individuals who evolve independently and may produce a variable number of offspring during their lifetime. MBTs have applications in biology, epidemiology and also in telecommunication systems. We refer the reader to [1, 4] for definitions, properties and applications.

One important issue related to MBTs is the computation of the extinction probability of the population, which can be characterized as the minimal nonnegative solution

\[ x^* \in \mathbb{R}_+^N, \quad \text{with } \mathbb{R}_+^N := \{ v \in \mathbb{R}^N : v_i \geq 0, i = 1, \ldots, N \}, \]

of the quadratic vector equation (QVE for short)

\[ x = a + b(x, x), \quad (1) \]

where \( a \in \mathbb{R}_+^N \) and \( b : \mathbb{R}_+^N \times \mathbb{R}_+^N \to \mathbb{R}_+^N \) is a vector-valued bilinear form such that the vector \( e = (1, 1, \ldots, 1)^T \) is always a solution of (1).

The MBT is called subcritical, supercritical or critical if the spectral radius \( \rho(R) \) of the matrix

\[ R := b(e, \cdot) + b(\cdot, e) \quad (2) \]

is strictly less than one, strictly greater than one, or equal to one, respectively.

Under the stated assumptions, one can prove the existence of a minimal nonnegative solution \( x^* \) in the componentwise ordering. A proof using minimal hypotheses is presented in [7]. In the subcritical and critical cases the minimal nonnegative solution is the vector of all ones, while in the supercritical case \( x^* \leq e, x^* \neq e \). Thus, only the supercritical case is of interest for the computation of \( x^* \).

The extinction probability \( x^* \) is usually computed by means of a functional iteration or Newton’s method applied directly to

\[ F(x) := a + b(x, x). \quad (3) \]

The sequence of approximations \( x_k \) resulting from this process is componentwise nondecreasing, and has a natural probabilistic interpretation; moreover, the applicability of the iteration and convergence to the correct solution is guaranteed [1, 4, 5]. However, when \( \rho(R) \) is close to 1 (which is, in the applications, the most interesting case), the convergence rate of the iterations degrades significantly. This is a behavior common to many similar matrix and vector equations, as we see in the following. Heuristically, this is due to the solution becoming close to a double solution, as \( x^* \approx e \).

In the case of this equation, however, one of these solutions is known \( (x = e) \); this suggest trying to perform a sort of deflation. Indeed, by setting \( y = e - x \), we may reduce the equation to the form

\[ y = H_y y, \quad (4) \]

where \( H_y = b(e - y, \cdot) - b(\cdot, e) \) is a matrix in \( \mathbb{R}_+^{N \times N} \). In other words, \( y \) is the Perron vector of a matrix depending linearly on \( y \) itself. We solve (4) by the fixed-point iteration

\[ y_{k+1} = \text{Perron}_\text{vector}(H_y), \]

The MBT is called subcritical, supercritical or critical if the spectral radius \( \rho(R) \) of the matrix

\[ R := b(e, \cdot) + b(\cdot, e) \]

is strictly less than one, strictly greater than one, or equal to one, respectively.
with a suitable normalization on $y_{k+1}$. We discuss implementation issues, variants (including Newton’s method) and convergence of this iteration. The interesting feature is that, while the customary methods slow down for close-to-critical problems, the convergence speed of this method increases. A limit-case analysis shows that for an exactly critical problem this method converges quadratically (instead of linearly), while the customary Newton method converges linearly (instead of quadratically). Numerical experiments show that on close-to-critical problems the Perron vector-based algorithm outperforms the customary ones [6].

Equation (1), with the positivity constraints on the parameters but without the hypothesis that $x = e$ is a solution, is general enough to encompass several matrix and vector equations that have been studied in literature; among them, the most important are the nonsymmetric algebraic Riccati equation $XCX + B = AX + XD$ [3], and the unilateral quadratic matrix equation used in the solution of QBD processes $X = PX^2 + QX + R$ [2]. The existing numerical methods for these equations show similar features (monotonic convergence, slowdown in close-to-critical cases), but up to our knowledge they have never been studied in a unified setting. It is possible carry out the main proofs in an unified fashion; in some cases, the hypotheses of the convergence theorems are relaxed. This generalization leads to a more precise understanding of the role of the (strict) positivity of the minimal solution and of the irreducibility of the Jacobian matrix of $F(x)$ in (3).

Since in this generalized case we do not have a solution $x = e$ known a priori, unfortunately it is not possible to apply the deflation strategy that lead to the Perron vector-based iteration. However, in the matrix equation case often partial information is known on the solution matrix, such as one of its eigenpairs. Our efforts concentrate now on finding a generalization of the Perron vector-based algorithms to this setting.

References


Symmetric Embeddings of Higher-Order Tensors

Stefan Ragnarsson and Charles F. Van Loan

Abstract

Well known connections exist between the singular value decomposition of a matrix $A$ and the Schur decomposition of its symmetric embedding $\text{sym}(A) = ([0 \ A ; A^T \ 0])$. In particular, if $\sigma$ is a singular value of $A$ then $+\sigma$ and $-\sigma$ are eigenvalues of the symmetric embedding. The top and bottom halves of $\text{sym}(A)$’s eigenvectors are singular vectors for $A$. Power methods applied to $A$ can be related to power methods applied to $\text{sym}(A)$. The rank of $\text{sym}(A)$ is twice the rank of $A$.

Here we show how to embed a general order-$d$ tensor $\mathcal{A}$ into an order-$d$ symmetric tensor $\text{sym}(\mathcal{A})$. Through the embedding we relate the tensor eigenvalues of $\text{sym}(\mathcal{A})$ to the tensor singular values of $\mathcal{A}$. We also use the embedding to connect power methods for symmetric tensors to power methods for general tensors and even derive new methods. The relationship between the rank of $\mathcal{A}$ to the rank of $\text{sym}(\mathcal{A})$ for both the multilinear rank and the outer product rank is also investigated.
Iterative Solution of Linear Systems in Liquid Crystal Modelling

Alison Ramage

Abstract

Although the mathematical theory of liquid crystals has been extensively studied for over 75 years, to date there has been much less work done on the numerous interesting and important numerical analysis issues which the study of such materials raises. Often, the underlying physical problems involve characteristic length and time scales which vary by many orders of magnitude, or complex combinations of fluid flow and changes in orientational order within a liquid crystal cell. Such features provide difficult numerical challenges to those trying to simulate the real-life dynamic situations which are of interest in an industrial setting. In particular, as is often the case with large PDE simulations, efficient solution of the resulting linear algebra subproblems is of crucial importance for the overall effectiveness of the algorithms used.

The focus of this work is on the iterative solution of saddle-point problems which occur frequently, and in multiple ways, in liquid crystal numerical modelling. For example, saddle-point problems arise whenever director models are implemented, through the use of Lagrange multipliers for the pointwise unit vector constraints (as opposed to using angle representations). In addition, saddle-point systems arise when an electric field is present that stems from a constant voltage, irrespective of whether a director model or a tensor model is used, or whether angle representations for directors or componentwise representation with pointwise unit vector constraints is employed. Furthermore, the combination of these two situations (that is, a director model using components, associated constraints and Lagrange multipliers, together with a coupled electric field interaction) results in a novel double/inner saddle-point structure which presents a particular challenge in terms of numerical linear algebra.

In this talk we will present some examples of the saddle-point systems which arise in liquid crystal modelling and discuss their efficient solution using appropriate preconditioned iterative methods with the aim of giving an insight into this new and exciting research area at the interface between liquid crystal theory and numerical analysis.
A Multipreconditioned GMRES Algorithm

_Tyrone Rees, Chen Greif, Daniel Szyld_

Abstract

Suppose we wish to solve the linear system $Ax = b$, where $A$ is a large, sparse square matrix. The solution method of choice is often a Krylov subspace method; e.g., the conjugate gradient method (CG) if $A$ is symmetric positive definite (s.p.d.), or GMRES for a general $A$. For such methods to be effective they invariably need to be applied along with a preconditioning matrix, chosen to improve the spectral properties of the system. The choice of a suitable preconditioner is in itself an art, and is highly problem-dependent. In many cases there are a number of suitable preconditioners, but with current algorithms the practitioner must pick just one.

In 2005 Bridson and Greif [1] expanded CG to a multipreconditioned conjugate gradient (MPCG) method, which allows more than one preconditioner. Like the underlying CG algorithm, this can be applied if we have a s.p.d. linear system and preconditioner. However, the attractive three term recurrence of CG is, in general, lost in the MPCG variant, which is a significant drawback of the algorithm.

One way of applying multiple preconditioners for a general matrix would be to use the flexible GMRES method [2] and cycle through the preconditioners, applying a different one at each step. This method would utilize the properties of all the preconditioners, but in a sub-optimal way. In this talk we present a multipreconditioned GMRES algorithm (MPGMRES) in which the preconditioners are automatically combined in an optimal way.

Given $t$ preconditioners, MPGMRES chooses a step length with minimizes the $\ell^2$-norm of the residual over the $t$ search directions at each iteration. The algorithm is implemented via a modified block Arnoldi procedure; when applied with a single preconditioner the method is simply the usual GMRES algorithm. The solution of the least squares problem requires $t$ Givens rotations at each iteration. This method can be applied to any matrix, and so unlike MPCG we are not restricted to s.p.d. matrices. Since the GMRES algorithm is not based on a short term recurrence we are not losing anything here by applying this method instead. MPGMRES can, like GMRES, be applied in a truncated or restarted version.

This algorithm is especially effective when the preconditioners act on different parts of the spectrum. The method has particular promise when solving saddle-point systems, as in certain cases it is possible to get a solver which does not require a good approximation to the Schur complement. We will present numerical results based on both idealized and practical linear systems to demonstrate the utility of the algorithm.

References


Recycling Krylov Subspaces for Efficient Schwarz Algorithms with Extensions to Solve Regularized Least Squares Problems

Youzuo Lin and Rosemary Renaut

Abstract

We consider the solution of the regularized least squares problem

\[
\min \| A x - b \|_W^2 + \| L x \|_W^2
\]

where \( A \in \mathbb{R}^{m \times n}, m \geq n \), is ill-conditioned, \( L \in \mathbb{R}^{p \times n}, p \leq n \) is a suitably chosen regularization operator, and at least one of the weighting matrices \( W_b, W_x \) is known. Note typically, \( W_b = I \) but statistical theory indicates that it may rather be the inverse covariance matrix for the noise in the measured data \( b \). While we may assume, without loss of generality, \( W_b = I \), and apply appropriate scaling to \( A \) and \( b \) by \( W_b^{1/2} = I \), which is trivial if the noise is Gaussian white or colored noise, it can also be useful in derivations involving determination of \( W_x \) to recall that the noise has been scaled to be standard normal. With respect to weighting on the model parameters, \( x, W_x \) is often taken to be a multiple of the identify, \( W_x = \lambda^2 I \), for some regularization parameter \( \lambda > 0 \). The research discussed here is more general, permitting regularization with parameter specific regularization parameters.

The method of multisplitting (MS) for finding the solution of linear systems of equations was introduced by O’Leary and White in 1985. If implemented through a straightforward column splitting of the matrix \( A \), when \( A \) is square, it amounts to a domain decomposition for the variables \( x \), and thus an option is implementation in additive or multiplicative Schwarz fashion (i.e. in the simplest form as Block Jacobi or Block Gauss Seidel). The extension of multisplitting for the solution of the full rank least squares problem was first considered by Renaut in 1998. Here we consider the solution of the ill-conditioned regularized problem through the domain decomposition - Schwarz model.

The novelty of the discussed approach is two fold - first the idea of recycling the Krylov subspace in conjunction with the domain decomposition appears to be new, and thus of more general interest than the specific problem (1), while second the capability to solve for the multiparameter problem efficiently opens the door for effectively determining the local weighting parameters which is a problem of significant current interest in the inverse problems community.

In this paper we first describe the standard MS algorithm as extended to solve (1). Basically the solution approach involves an inner-outer iteration strategy in which the inner steps involve parallel solution of local problems and the outer step is the global update of the solution. For each global outer iteration, solutions of the local problems \( A_i y_i \approx b_i(x^{(k)}) \), for each \( i \) are required. The system matrix \( A_i \) is fixed for all inner steps, but the right hand depends on the most recent update \( x^{(k)} \) at step \( k \). Notice for the regularized problem (1) matrix \( A_i \) is the subproblem augmented matrix for (1), and \( b_i(x^{(k)}) \) is defined accordingly. There is a significant literature on the use of Krylov subspace recycling for systems with multiple right hand sides. This problem falls into that category, but within the framework for right hand sides that are nearby. Notice that for a convergent algorithm as \( x^{(k)} \) tends to \( x^* \), the solution of (1), \( b(x^{(k)}) \) tends to \( b(x^*) \). In particular, provided the conditioning of the problem has been handled appropriately by the regularization term, changes in the right hand side decrease with the iteration. It is thus appropriate to recycle the seed Krylov subspace, that obtained at the first outer iteration \( k = 0 \), for subsequent steps. We
describe two approaches, one in which that initial space is augmented as $k$ increases, and a second in which a restart is introduced.

Theoretical and experimental results concerning the recycling approach in the context of the solution of (1) are presented, with practical details such as tolerance, reorthogonalization and a GPU implementation also addressed. We then anticipate showing results which utilize the framework for solution with a multiparameter estimation included.
New Methods for Least-Norm Regularization

Danny C. Sorensen, Marielba Rojas

Abstract
We consider the least-norm regularization problem

\[ \min \|x\| \quad \text{subject to} \quad \|Ax - b\| \leq \epsilon \]  

where \( A \) is an \( m \times n \) real matrix, \( b \) is an \( m \)-dimensional real vector, \( \epsilon \) is a positive scalar, and \( \| \cdot \| \) is the Euclidean norm. Problem (1) arises in the regularization of discrete forms of ill-posed problems where \( A \) is a discretized version of an operator in an ill-posed problem, and \( b \) is a data vector contaminated by noise. In particular, we assume \( b = b_o + n \) where \( b_o \) is the noise-free, unknown data vector such that \( b_o = Ax_o \) with \( x_o \) the desired solution, and where \( n \) represents noise. We also assume that \( \epsilon \) is an upper bound on the size of the noise, ie. \( \|n\| \leq \epsilon \).

Problem (1) was first studied in [2, 4, 8]. Solution methods for (1) were proposed in [2, 7, 8, 11]. Methods for a related problem were proposed in [1, 3, 5].

We derive optimality conditions for problem (1) and establish new conditions for feasibility. We present two classes of methods: a factorization-based algorithm for solving the associated secular equation in the dense case, and two matrix-free iterations for large-scale problems.

The dense algorithm is a robust implementation of Newton’s method that incorporates the ideas in [6] and uses the new initial estimate proposed in [10]. The matrix-free iterations are iterative projection methods inspired by Voss’s Nonlinear Arnoldi/Lanczos method [12] and based on a suitable reformulation of the optimality conditions. The new methods use the dense algorithm to solve secular equations for small, projected problems.

We derive a residual-space iteration, geared towards the case \( m \ll n \), and a solution-space iteration, geared towards the case \( m \gg n \). At each step of these iterations, an existing basis is expanded. A linear Lanczos method would expand this basis with a vector obtained by the 3-term Lanczos recurrence and this would amount to a standard Krylov update. In contrast, our iterations obtain the new expansion vector from a nonlinear process, namely, the solution of the secular equation. The new vector is either in the direction of a projected gradient, or a Newton-like step if a preconditioner is used. In many ways, this is analogous to the expansion step in a Jacobi-Davidson scheme [9].

An important feature of the new methods is the possibility of incorporating both fixed and variable preconditioning in a straightforward manner.

We present numerical results that demonstrate that the new methods are accurate, efficient, and robust.
References

Orthogonalization With a Non-standard Inner Product and Approximate Inverse Preconditioning

Jiří Kopal, Miro Rozložník, Miroslav Tůma and Alicja Smoktunowicz

Abstract

One of the most important and frequently used preconditioning techniques for solving symmetric positive definite systems $Ax = b$ is based on computing the approximate inverse factorization in the form $A^{-1} = ZZ^T$, where $Z$ is upper triangular [1]. It is also a well-known fact that the factor $Z$ can be computed column by column by means of an $A$-orthogonalization process applied to the unit basis vectors $e_1, \ldots, e_n$. As noted in [3] such $A$-orthogonalization also produces the Cholesky factor of the matrix $A = U^T U$ where $U^{-1} = Z$. This fact has been exploited to construct efficient sparse approximate inverse factorization preconditioners [1, 2, 3]. In a more general setting, given the symmetric positive definite matrix $A$ and the nonsingular matrix $Z^{(0)}$, we look for the factors $Z$ and $U$ so that $Z^{(0)} = ZU$ with $Z^T AZ = I$ and the upper triangular $U$ is a Cholesky factor of the matrix $(Z^{(0)})^T AZ^{(0)} = U^T U$.

There are several ways how to compute the matrices $Z$ and $U$. If we have the spectral decomposition $A = V \Lambda V^T$, the factor $U$ can be obtained from the standard QR decomposition $\Lambda^{1/2} V^T Z^{(0)} = QU$. The factor $Z$ can be then recovered as $Z = V \Lambda^{-1/2} Q$. Probably the most straightforward and frequently used approach is the Gram-Schmidt orthogonalization, which consecutively $A$-orthogonalizes the columns of $Z^{(0)}$ against previously computed vectors using the orthogonalization coefficients that form the factor $U$. In the classical Gram-Schmidt algorithm (CGS), the $A$-orthogonal vectors are computed via matrix-vector updates which are relatively easy to parallelize. The rearrangement of this scheme has led to the modified Gram-Schmidt algorithm (MGS) with better numerical properties. Introducing sequential orthogonalization however destroys desirable parallel properties of the algorithm. We will discuss also yet another variant of sequential orthogonalization, which is motivated originally by the AINV preconditioner and which uses oblique projections [3]. We will refer to this scheme as the AINV orthogonalization. The main motivation for the development of approximate inverse techniques came from parallel processing and so the early papers on inverse factorization did not study numerical properties of algorithms. However, concerns on robustness and accuracy became very quickly an important aspect and resulted into a significant progress in recent preconditioning techniques. While the initial schemes like the basic AINV algorithm were based on oblique projections or on the CGS orthogonalization [3], the development lead to their stabilization both in terms of the orthogonalization scheme (MGS in the SAINV algorithm [2]) and in terms of appropriate computation of diagonal entries in $U$ (one-sided versus stabilized versions of AINV [2, 3]).

From a numerical point of view, all these techniques may produce vectors which are far from orthogonal. The orthogonality between computed vectors is however crucial for the quality of the preconditioner constructed in the approximate inverse factorization. Given some approximation $\tilde{Z}$ to $Z$ such that $A^{-1} \approx \tilde{Z} \tilde{Z}^T$, we are especially interested in the loss of orthogonality between the columns of $\tilde{Z}$ measured by the 2-norm of the matrix $\tilde{Z}^T A \tilde{Z} - I$. It is a well-known fact that eigenvalues of $\tilde{Z}^T A \tilde{Z}$ determine the convergence rate of the preconditioned conjugate gradient method applied to $\tilde{Z}^T A \tilde{Z} y = \tilde{Z}^T b$ where $x = \tilde{Z} y$. Therefore the orthogonal basis problem is of a primary interest for this application. While for the case of the standard inner product there exists a complete rounding error analysis for all main orthogonalization schemes [4, 7], the numerical properties of the schemes with a non-standard inner product are much less understood.
In this contribution we review the most important schemes used for orthogonalization with respect to the non-standard inner product and give the worst-case bounds for corresponding quantities computed in finite precision arithmetic. We formulate our results on the loss of orthogonality and on the factorization error (measured by \( \| \bar{Z}^T A Z - I \| \) and \( \| Z^{(0)} - \bar{Z} U \| \)) in terms of quantities proportional to the roundoff unit \( u \), in terms of the condition number \( \kappa(A) \) which represents an upper bound for the relative error in computing the \( A \)-inner product as well as the condition number of the matrix \( A^{1/2} Z^{(0)} \) which plays an important role in the factorization \( (Z^{(0)})^T A Z^{(0)} \approx \bar{U}^T \bar{U} \).

Although all orthogonalization schemes are mathematically equivalent, their numerical behavior can be significantly different. It follows from our analysis that while the factorization error is quite comparable for all these schemes, the orthogonality between computed vectors can be significantly lost and it depends on the condition number \( \kappa(A) \). This is the case also for the eigenvalue-based (EIG) implementation and Gram-Schmidt with reorthogonalization (CGS2). The classical Gram-Schmidt algorithm and AINV orthogonalization behave very similarly and generate vectors with the orthogonality that besides \( \kappa(A) \) depends also on the factor \( \kappa(A^{1/2} Z^{(0)}) \kappa(Z^{(0)}) \) (it essentially means the quadratic dependence on the condition number of the matrix \( A^{1/2} Z^{(0)} \)). Since the orthogonality in the modified Gram-Schmidt algorithm depends only linearly on \( \kappa(A^{1/2} Z^{(0)}) \), MGS appears to be a good compromise between expensive EIG or CGS2 and less accurate CGS or AINV. Indeed in the context of approximate inverse preconditioning the stabilization of AINV has lead to the SAINV algorithm which uses exactly the MGS orthogonalization. We treat separately the particular case of a diagonal \( A \) which is extremely useful in the context of weighted least squares problems. One can show then that local errors arising in the computation of a non-standard inner product do not play an important role here and that the numerical behavior of these schemes is almost identical to the behavior of the orthogonalization schemes with the standard inner product. For all these results we refer to [6]. We believe that these results are an initial step towards understanding the behavior of practical strategies in approximate inverse preconditioning and will stimulate further research of schemes that use some inexact orthogonalization with appropriate dropping criterion and lead to some sparse approximation of \( Z \) and \( U \). For a survey of such preconditioning techniques we refer to [1].

References

Recent Developments of Rational Krylov Algorithms

Axel Ruhe

Abstract

The Rational Krylov eigenvalue algorithm has got a rather wide spread use in model reduction for linear dynamic systems. We will study some implementation issues, first on how to make sure that structural properties of the original system are kept in the reduced model, a real system should have a real reduced model, then the choice of shifts and poles, the transfer function of the reduced system will be an interpolation to the original transfer function.
A New Approach to GMRES Convergence

Mohamed Bellalij Hassane Sadok

Abstract

GMRES is one of the most widely used iterative methods for the solution of linear system of equations

\[ Ax = b, \]

with a large real or complex nonsingular \( n \times n \) matrix \( A \) and a real or complex right-hand side \( b \); see, e.g., [6, 7]. Let \( x_0 \) be an available approximate solution of (1), define the associated residual vector \( r_0 = b - Ax_0 \), and introduce the Krylov subspace \( K_k(A, r_0) = \text{span}\{r_0, Ar_0, \ldots, A^{k-1}r_0\} \).

The \( k \)th iterate, \( x_k \), determined by GMRES lives in \( K_k(A, r_0) \) and is chosen so that the associated residual error \( r_k = b - Ax_k \) satisfies

\[ \|r_k\| = \min_{z \in K_k(A, r_0)} \|b - A(x_0 + z)\|; \]

see, e.g., [6]. Here and throughout this paper \( \| \cdot \| \) denotes the Euclidean vector norm or the associated induced matrix norm.

Convergence properties of GMRES are discussed, e.g., in [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. Most convergence results are obtained by expressing (2) as a polynomial approximation problem. Let \( P_k \) denote the set of polynomials of degree at most \( k \) and let \( P_k^{(0)} \) denote the subset of these polynomials \( p \) with \( p(0) = 1 \). Then it follows from (2) that \( r_k = p_k(A) r_0 \) for some \( p_k \in P_k^{(0)} \). It is straightforward to bound \( \|r_k\| \) under the assumption that \( A \) is diagonalizable, i.e.,

\[ A = S \Lambda S^{-1}, \quad S = [s_1, \ldots, s_n], \quad \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n). \]

Unless explicitly stated otherwise, the columns \( s_j \) of \( S \) are assumed to be of unit Euclidean norm. We have \( r_k = S p_k(\Lambda) S^{-1}r_0 \) and therefore

\[ \|r_k\| \leq \|r_0\| \|S\| \|S^{-1}\| \min_{p \in P_k^{(0)}} \max_{i=1,\ldots,n} |p(\lambda_i)|, \]

see, e.g., [6]. Here and throughout this paper \( \| \cdot \| \) denotes the Euclidean vector norm or the associated induced matrix norm.

If in addition \( A \) is normal, then we may choose the matrix \( S \) to be orthogonal and the bound (4) simplifies to

\[ \|r_k\| \leq \|r_0\| \min_{p \in P_k^{(0)}} \max_{i=1,\ldots,n} |p(\lambda_i)|. \]

This bound is sharp in the sense that for any \( k \geq 1 \), there is a normal matrix \( A \) and an initial residual vector \( r_0 \), such that equality is achieved. However, we will show that if equality holds in (5) for some \( k = k < n \), then the iterate \( x_{k+1} \) solves (1). Thus, the bound (5) is not sharp for the common situation that the GMRES iterations do not terminate early.

It is the aim of the present paper to express the convergence of the norm of the residual errors \( \|r_k\| \) as \( k \) increases in terms of the eigenvalues of \( A \) and the initial residual vector \( r_0 \). Section 2 defines GMRES in using Petrov-Galerkin conditions. The norm of the computed residual, \( \|r_k\| \), then is expressed with the aid of the inverse of the matrix \( K_k^H K_k \), where

\[ K_k = [r_0, Ar_0, \ldots, A^{k-1}r_0] \]

is the Krylov matrix and the superscript \( ^H \) denotes transposition and complex conjugation.
Bounding the norm of the residual vectors $r_k$ determined by GMRES in terms of the eigenvalues of the matrix $A$, when $A$ is a general square nonsingular matrix, is a difficult constrained optimization problem. We therefore in the sequel of this paper focus on diagonalizable matrices. Section 3 relates the norm of the residual vectors $r_k$ associated with GMRES iterates for a linear system (1) with a diagonalizable matrix to the norm of residual vectors obtained when GMRES is applied to linear system of equations with an associated normal matrix and a particular initial iterate. The application of GMRES to linear systems of equations with a normal matrix is discussed in Section 4. There we first present an explicit formula for $\|r_1\|$ in terms of the eigenvalues of $A$, and then bound $\|r_k\|$ for $k > 1$ by solving a constrained optimization problem using Lagrange multipliers.

This paper complements previous investigations by Greenbaum [1], Ipsen [3], and Liesen and Tichy [5]. Greenbaum [1] derived related bounds for the situation when the matrix $A$ is symmetric positive definite. Ipsen’s investigation [3] is closely related to the one of the present paper. She derived a different optimization problem, which leads to a bound [3, Theorem 3.1] for $\|r_k\|$ different from ours. In particular, the bound of [3, Theorem 3.1] contains a constant $c_{i+1}$, whose exact value is difficult to determine from the eigenvalues. Finally, Liesen and Tichy [5] obtain bounds similar to ours for the norm of the penultimate residual error.

References


AMRES: The Minimum-Residual Method for Augmented Systems

David Fong, Per Christian Hansen, and Michael Saunders

Abstract
By specializing the iterative solver MINRES [1], we derive a solver for augmented systems of the form
\[
\begin{pmatrix}
\gamma I & A \\
A^T & \delta I
\end{pmatrix}
\begin{pmatrix}
s \\
x
\end{pmatrix}
= 
\begin{pmatrix}
b \\
0
\end{pmatrix},
\]

where \( A \) is a rectangular sparse matrix and \( \gamma, \delta \) are given scalars with \( \gamma \neq 0 \). If \( \delta = -\gamma \), the system is equivalent to the damped least-squares problem \((AT^2I)x = ATb\), and it would be efficient to apply CGLS, LSQR, or LSMR [2, 3, 5]. If \( \delta = \gamma \), the system is equivalent to the negatively damped problem \((AT^2I)x = ATb\) that arises in total least squares problems as well as in certain problems in \( H^\infty \) control. Our solver AMRES provides a stable method for this case.

If \((\sigma, u, v)\) is one of the singular value triplets of \( A \), we know that \( \gamma = \delta = -\sigma \) gives a singular system (and it is incompatible if \( ATb \neq 0 \)). We could compute the singular vector \( v \) in the manner of inverse iteration by applying AMRES with a random \( b \), allowing the iterates \( \{x_k\} \) to diverge, and taking \( v = x_k/\|x_k\| \) when the norm becomes very large. Alternatively, following the approach described in [4] for computing nullvectors, we could regard AMRES as solving the singular least-squares problem
\[
\min_{s, x} \left\| \begin{pmatrix}
-\sigma I & A \\
A^T & -\sigma I
\end{pmatrix}
\begin{pmatrix}
s \\
x
\end{pmatrix}
- 
\begin{pmatrix}
b \\
0
\end{pmatrix} \right\|.
\]
The iterates will converge to bounded values \((s_\ell, x_\ell)\) before diverging. At this point, the residual
\[
\begin{pmatrix}
u_\ell \\
v_\ell
\end{pmatrix}
= 
\begin{pmatrix}
b \\
0
\end{pmatrix}
- 
\begin{pmatrix}
-\sigma I & A \\
A^T & -\sigma I
\end{pmatrix}
\begin{pmatrix}
s_\ell \\
x_\ell
\end{pmatrix}
\]
should give (unnormalized) approximate singular vectors. We will describe the AMRES implementation (based on Golub-Kahan bidiagonalization of \( A \)) and its use for computing singular vectors in this way when an accurate singular value \( \sigma \) is known.

Multilinear Rank-\((r_1, r_2, r_3)\) Approximation of a Tensor: Optimality Conditions and Perturbation Theory of Local Optima

Berkant Savas and Lars Eldén

Abstract

In recent years multilinear algebra has gained considerable attention in various research fields [6]. One contributing factor for this is the increase of applications where the data naturally has a multilinear structure. The objects of interests in numerical multilinear algebra are tensors. An order-3 tensor \(A\) may be considered as an array with three indices, i.e., \(A = [a_{ijk}]_{i,j,k=1}^{l,m,n} \in \mathbb{R}^{l \times m \times n}\), with associated algebraic operations, for example addition, multiplication and various contractions. Thus, tensors are generalizations of vectors and matrices, which are order-1 and order-2 tensors, respectively.

We consider the problem of approximating a given tensor \(A \in \mathbb{R}^{l \times m \times n}\) with another tensor \(B\) that has multilinear rank-\((r_1, r_2, r_3)\), i.e.,

\[
\min_{B} \|A - B\|_F^2,
\]

under the constraint \(\text{rank}(B) \leq (r_1, r_2, r_3)\). Any tensor \(B \in \mathbb{R}^{l \times m \times n}\) with \(\text{rank}(B) = (r_1, r_2, r_3)\) can be written in factorized form using some matrices \(X \in \mathbb{R}^{l \times r_1}\), \(Y \in \mathbb{R}^{m \times r_2}\), \(Z \in \mathbb{R}^{n \times r_3}\) and a small core tensor \(C \in \mathbb{R}^{r_1 \times r_2 \times r_3}\),

\[
B = (X, Y, Z) \cdot C,
\]

where \(b_{ijk} = \sum_{\lambda=1}^{r_1} \sum_{\mu=1}^{r_2} \sum_{\nu=1}^{r_3} x_{i\lambda} y_{j\mu} z_{k\nu} c_{\lambda\mu\nu}\).

This operation defines the multilinear product between a tensor and a set of matrices. Using the factored form \(B = (X, Y, Z) \cdot C\) we can write the approximation problem as

\[
\min_{X, Y, Z, C} \|A - (X, Y, Z) \cdot C\|_F^2
\]

which can be shown to be equivalent with the maximization problem

\[
\max_{X, Y, Z, C} \Phi(X, Y, Z), \quad \text{where} \quad \Phi(X, Y, Z) = \|A \cdot (X, Y, Z)\|_F^2,
\]

with orthogonality constraints \(X^T X = I\), \(Y^T Y = I\), \(Z^T Z = I\) [4]. Observe that the small core tensor \(C\) is not present in the maximization problem. Once \(X\), \(Y\), and \(Z\) are determined the core tensor is uniquely determined, namely \(C = A \cdot (X, Y, Z)\). Writing out the objective function \(\Phi(X, Y, Z)\) reveals that it is a sixth degree polynomial in the entries of \(X\), \(Y\) and \(Z\). Clearly the maximization problem and therefore also the tensor approximation problem is highly nonlinear. Given a point \((X, Y, Z)\), that yields a local maximum of \(\Phi(X, Y, Z)\), there is no (simple and practical) way of determining whether it is the global maximum that would give the best low rank approximation. The tensor approximation problem has many local minima, and the “best” low multilinear rank approximation of a tensor attempts to find local maximum points to \(\Phi(X, Y, Z)\) [4, 7, 3].

In this talk we will present first and second order optimality conditions of a local maximizer of \(\Phi(X, Y, Z)\). A preliminary analysis of the first order case is presented in [1]. The first and second order optimality conditions have direct relations to the structure of the problem. We will show what this structure and these relations are, and compare with the corresponding structure for the low rank matrix approximation problem. The matrix approximation problem is well understood, it is
convex and the global optimizer is readily obtained by the truncated singular value decomposition (SVD) [5]. We will show that the structure of the optimality conditions closely resemble the all-orthogonality and ordering properties in the Higher Order SVD [2] of a given tensor.

A second topic of this talk is perturbation theory. It is often important to know how sensitive the low rank approximation of a matrix is with respect to perturbations. For a given matrix $A \in \mathbb{R}^{m \times n}$, consider its SVD: $A = U \Sigma V^T$, and its best rank-$k$ approximation: $U_1 \Sigma_1 V_1^T \approx A$, where we have partitioned $U = [U_1 \ U_2]$, $V = [V_1 \ V_2]$ and $\Sigma = \text{diag}(\Sigma_1, \Sigma_2)$. $U_1$ and $V_1$ contain the leading $k$ left and right singular vectors, respectively, and $\Sigma_1$ contains the corresponding singular values. It is well known that sensitivity of $U_1$ and $V_1$ is bounded in terms of the “gap” between $\sigma_k$ and $\sigma_{k+1}$ [8]. The larger $\sigma_k - \sigma_{k+1}$ the less sensitive are $U_1$ and $V_1$ to perturbations. Perturbation analysis for the tensor case is more involved, but we will show that the concept of gap can be generalized to tensors. For the matrix case, the gap between $\sigma_k$ and $\sigma_{k+1}$ can also be interpreted in terms of $\Sigma_1 = U_1^T A V_1$ and $\Sigma_2 = U_2^T A V_2$. Consider the special case with rank-$(1, 1, 1)$ approximation of a tensor and let a local optimum be given by the three vectors $(x, y, z)$. Then the sensitivity depends on (roughly) the relative sizes of $A \cdot (x, y, z)$ (which corresponds to the $\Sigma_1$ term from the matrix case) and the three matrices $A \cdot (X_\perp, Y_\perp, x)$, $A \cdot (X_\perp, y, Z_\perp)$ and $A \cdot (x, Y_\perp, Z_\perp)$ (all of which are related to the $\Sigma_2$ term from the matrix case). Note that we have introduced orthogonal complements $X_\perp$, $Y_\perp$, and $Z_\perp$ so that $[x \ X_\perp]$, $[y \ Y_\perp]$, $[z \ Z_\perp]$ form orthogonal matrices. For general rank-$(r_1, r_2, r_3)$ approximation additional terms are involved in determining the sensitivity. We will describe all the terms that are involved in the equations for the perturbations. Finally, we will present a few numerical experiments illustrating the presented analysis of the perturbation.

References


A Jacobi-Davidson Method for Two Real Parameter Nonlinear Eigenvalue Problems arising from Delay Differential Equations

Karl Meerbergen, Christian Schröder, and Heinrich Voss

Abstract

Consider the time-invariant linear delay-differential equation (DDE)

\[ M \dot{x}(t) + Ax(t) + Bx(t - \tau) = 0 \]  

where, \( M, A, B \in \mathbb{C}^{n \times n} \) are given system matrices and \( \tau \geq 0 \) is the delay. This system is (asymptotically) stable if, for every bounded initial condition, it holds that \( x(t) \to 0 \) as \( t \to \infty \), and a necessary and sufficient condition is that the spectrum of the eigenvalue problem

\[ \lambda Mu + Au + e^{-\tau \lambda} Bu = 0 \]

is contained in the open left half–plane.

Our goal is to find the delays \( \tau \) for which the DDE (1) is stable.

Due to the continuous dependence of eigenvalues on the parameter \( \tau \), the stability behavior of (1) can only change as the delay \( \tau \) varies if an eigenvalue of (2) crosses the imaginary axis. Hence, our goal is reached by finding real parameters \( \tau \geq 0 \) called critical (or switching or crossing) delays such that \( \lambda = i\omega, \omega \in \mathbb{R} \) is an eigenvalue of (2), i.e. such that the parameter dependent delay eigenvalue problem

\[ T(\omega, \tau)u := i\omega Mu + Au + e^{-i\omega \tau} Bu = 0 \]

has a real eigenvalue \( \omega \).

Equation (3) is a specific case of the two real parameter eigenvalue problem which is to find real values \( \omega, \tau \in \mathbb{R} \) and a complex nonzero vector \( u \in \mathbb{C}^n \setminus \{0\} \) such that

\[ T(\omega, \tau)u = 0 \]

for a given map \( T : \mathbb{R} \times \mathbb{R} \to \mathbb{C}^{n \times n} \). Other examples arise from the stability analysis of hydrodynamic systems and chemical reactors, where the parameter \( \tau \) is the Reynolds number or the temperature, respectively. The obvious approach of continuation (alias homotopy) was shown in the literature to be inefficient and possibly leading to wrong conclusions.

Our approach depends on the dimension \( n \) of the system (1). For small scale problems we derive a real matrix pencil of dimension \( 2n^2 \) such that the eigentriples of (3) with real \( \omega, \tau \) can be reconstructed from the real eigenvalues and corresponding eigenvectors of this pencil. Since the real eigenvalues of a real pencil can be determined in a stable way by the real QZ algorithm, the wanted eigentriples of (3), and therefore the critical delays can be identified in a reliable way.

Unfortunately the complexity of this approach is \( O(n^6) \) such that it only applies to relatively small dimensions. To deal with medium sized to large scale problems, we combine the structure preserving eigensolver with an iterative projection method of Jacobi-Davidson type. Let \( V \in \mathbb{C}^{n \times k}, k \ll n \) be an orthonormal basis of an ansatz space, and let \( (\hat{\omega}, \hat{\tau}, \hat{y}), \hat{\omega}, \hat{\tau} \in \mathbb{R} \) be a solution of the projected \( k \) dimensional eigenvalue problem

\[ V^H T(\omega, \tau)V y = 0 \]

which has the same structure as the original problem but can be solved by above structure preserving method. If the resulting approximate solution \( (\hat{\omega}, \hat{\tau}, \hat{u}) \) of (3) with \( \hat{u} = V \hat{y} \) does not satisfy our
accuracy requirements then $V$ is expanded following an analogous goal as in the Jacobi–Davidson method for nonlinear eigenvalue problems: (an approximation to) the Newton direction of (3) has to be contained in the new search space $\text{span}[V, c]$.

More precisely, we are looking for a correction $(\hat{\omega} + \delta, \hat{\tau} + \varepsilon, \hat{u} + c)$ where $\delta, \varepsilon$ are real and $c$ is complex and is orthogonal to $\hat{u}$. This orthogonality constraint makes the direction of $c$ unique. Applying a single step of Newtons method to (3) leads to the Newton correction equation

$$
\begin{bmatrix}
T(\hat{\omega}, \hat{\tau}) & T_\omega(\hat{\omega}, \hat{\tau}) \hat{u} & T_\tau(\hat{\omega}, \hat{\tau}) \hat{u} \\
\hat{u}^H & 0 & 0 \\
0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
c \\
\delta \\
\varepsilon
\end{bmatrix}
= 
\begin{bmatrix}
-r \\
0 \\
0
\end{bmatrix}
:= 
\begin{bmatrix}
-T(\hat{\omega}, \hat{\tau}) \hat{u}
\end{bmatrix},
\tag{4}
$$

which is a linear system of $(n+1)$ complex equations in $n$ complex and 2 real unknowns. For medium scale dimensions $n$ we construct an equivalent linear system with $2n + 2$ equations and unknowns, which can be tackled by (sparse) direct solvers.

For truly large dimensions $n$, the linear system has to be solved by an iterative solver. We transform (4) to the following form which is more suitable for a Krylov solver:

$$
(I - K(U^H K)^{-1} U^H) T (I - U U^H) \begin{bmatrix}
c \\
\delta \\
\varepsilon
\end{bmatrix}
= 
-(I - K(U^H K)^{-1} U^H) \begin{bmatrix}
r \\
r \\
r
\end{bmatrix}
\tag{5}
$$

where $T, K, U$ are constructed from $T, \hat{u}, \hat{\omega}, \hat{\tau}$. Equation (5) looks like the common correction equation of the Jacobi–Davidson method for nonlinear eigenvalue problems: the equation

$$
T \begin{bmatrix}
c \\
\delta \\
\varepsilon
\end{bmatrix}
= 
\begin{bmatrix}
r \\
r \\
r
\end{bmatrix}
$$

is complemented by the oblique projector $I - K(U^H K)^{-1} U^H$ which provokes that the Newton direction is contained in $\text{span}\{\hat{u}, c\}$. We call (5) the JD correction equation.

For Krylov solvers to work efficiently, the use of a preconditioner is highly recommended. We derive a preconditioner $\tilde{P}$ for (5) from a preconditioner $P$ for $T(\hat{\omega}, \hat{\tau})$. Then each iteration of the Krylov solver for (5) requires only one matrix-vector product $T(\hat{\omega}, \hat{\tau}) v$ and one solve with the preconditioner $P$, and the only additional cost taking into account the projectors in the preconditioning are two matrix-vector products and three solves with the preconditioner $P$ to initialize the Krylov solver. This favorable property that the inclusion of the projectors into the preconditioner comes nearly for free was already pointed out for the original Jacobi Davidson method.

To illustrate its performance the method is applied to delay equations arising from discretized PDEs, demonstrating that the method is very robust with respect to inexact solves of the correction equation, coarse preconditioners, and the starting vector.

Details can be found in [1].

References

Graph Expansion and Communication Costs of Fast Matrix Multiplication

Grey Ballard¹, James Demmel², Olga Holtz³, Oded Schwartz⁴

Abstract

The communication of an algorithm (e.g., transferring data between the CPU and memory devices, or between parallel processors, a.k.a. I/O-complexity) often costs significantly more time than its arithmetic. It is therefore of interest to design and implement algorithms minimizing communication on the one hand, and to obtain lower bounds for the communication needed, on the other hand.

Previous Work. In [2, 3] we generalize the results of [10, 12] regarding matrix multiplication, to attain new I/O-complexity lower bounds for a much wider variety of algorithms (most of the bounds were shown to be tight). This includes algorithms for LU factorization, Cholesky factorization, LDLᵀ factorization, QR factorization, as well as algorithms for eigenvalues and singular values. Thus we essentially cover all direct methods of linear algebra. The results hold for dense matrix algorithms (most of them are of cubic time), as well as sparse matrix algorithms (whose running time depends on the number of non-zero elements). They apply to sequential and parallel algorithms, to compositions of linear algebra operations (like computing the powers of a matrix), and to certain graph theoretic problems⁵.

In [2, 3] we use the approach of [12], based on the Loomis-Whitney geometric theorem [13, 4], by embedding segments of the computation process into a three dimensional cube. This approach, however, is not suitable when distributivity is used, as is the case in Strassen [16] and other fast matrix-multiplication algorithms (e.g., [7, 6]).

The I/O-complexity of classic matrix multiplication and algorithms with similar structure is quite well understood. This is not the case for algorithms of more complex structure. Avoiding the communication of parallel classical matrix multiplication was addressed [5] almost simultaneously with the publication of Strassen’s fast matrix-multiplication [16]. Moreover, an I/O-complexity lower bound for the classical matrix-multiplication algorithm is known for almost three decades. Nevertheless, the I/O-complexity of Strassen’s fast matrix multiplication and similar algorithms has not yet been resolved.

Communication Cost of Fast Matrix Multiplication.

Upper bound. The I/O-complexity IO(n) of Strassen’s algorithm applied to n-by-n matrices on a machine with fast memory of size M, can be bounded above as follows (for actual uses of Strassen’s algorithm, see [9, 11, 8]): Run the recursion until matrices are sufficiently small. Then, read the two input matrices into the fast memory, perform the matrix multiplication inside the fast memory, and write the result into the slow memory. We thus have IO(n) ≤ 7 · IO (n/2) + O(n²) and

\[ IO \left( \frac{\sqrt{M}}{2} \right) = O(M). \]

Thus

\[ IO(n) = O \left( \left( \frac{n}{\sqrt{M}} \right)^{\omega_0} \cdot M \right). \]  

¹Computer Science Division, University of California-Berkeley
²Department of Mathematics and Computer Science Division, University of California-Berkeley
³Departments of Mathematics, University of California-Berkeley and Technische Universität Berlin
⁴Weizmann Institute of Science, ISRAEL
⁵See [14] for bounds on graph-related problems, and our [3] for a detailed list of previously known and recently designed sequential and parallel algorithms that attain the above mentioned lower bounds.
where $\omega_0 = \lg 7$.

**Lower bound.** In this paper, we obtain a tight lower bound:

**Theorem 1.** (Main Theorem) The I/O-complexity $IO(n)$ of Strassen’s algorithm on a machine with fast memory of size $M$ is

$$IO(n) = \Omega \left( \left( \frac{n}{\sqrt{M}} \right)^{\omega_0} \cdot M \right).$$  \hspace{1cm} (2)

It holds for any implementation and any known variant of Strassen’s algorithm. This includes Winograd’s $O(n^{\omega_0})$ variant that uses 15 additions instead of 18, which is the most used fast matrix multiplication algorithm in practice [9].

We then extend the lower-bound to a wider class of all stationary (i.e., Strassen-like) and non-stationary but uniform fast matrix-multiplication algorithms, and obtain tight lower bounds for these algorithms as well (where the new value of $w_0$ corresponds to the relevant algorithm). We also conclude a corresponding lower bound for parallel implementation of these algorithm.

**The Expansion Approach.** The proof of the main theorem is similar to the one taken by Hong and Kung [10] and is based on estimating the edge expansion of the computation directed acyclic graph (CDAG) of an algorithm (where we have a vertex for each input / intermediate / output argument, and edges according to dependencies).

An implementation of an algorithm determines, in the parallel model, which arithmetic operations are performed by which of the $p$ processor. This corresponds to partitioning the corresponding CDAG into $p$ parts. Edges crossing between the various parts, correspond to arguments that are in the possession of one processor, but are needed by other processor, and therefore relate to communication. A corresponding interpretation can be given for the sequential model.

The I/O-complexity is thus tightly connected to the edge expansion properties of this graph. As the graph has a recursive structure, the expansion can be analyzed directly (combinatorially, similarly to what is done in [1]) or by spectral analysis (in the spirit of what was done for the Zig-Zag expanders [15]). There is however, a new technical challenge. While the replacement and Zig-Zag products act similarly on all vertices, this is not the case here: multiplication and addition vertices behave differently when applying a recursive step.

**References**


The Robust and Efficient Partial Factorization of Dense Symmetric Indefinite Matrices

John Reid and Jennifer Scott

Abstract
Symmetric indefinite linear systems of equations arise in a wide range of applications including least-squares problems, optimization, fluid flow, electromagnetic scattering and eigenvalue problems. Our primary interest is in large sparse indefinite systems. One popular and well-established approach for solving such systems is the multifrontal method. At the heart of the multifrontal method is the need to efficiently and stably factorize dense indefinite \( n \times n \) matrices of the form

\[
A = \begin{pmatrix}
A_{11} & A_{21}^T \\
A_{21} & A_{22}
\end{pmatrix},
\]

where \( A_{11} \) is a square matrix of order \( p \leq n \) and pivots are restricted to lying within \( A_{11} \). Our factorization takes the form

\[
A = Q \begin{pmatrix}
L_1 & 0 \\
L_2 & I
\end{pmatrix} \begin{pmatrix}
D_1 & 0 & 0 \\
0 & A_S & 0 \\
0 & 0 & I
\end{pmatrix} \begin{pmatrix}
L_1^T & L_2^T \\
0 & I
\end{pmatrix} Q^T
\]

where \( Q \) is a permutation matrix

\[
Q = \begin{pmatrix}
Q_1 & 0 \\
0 & I
\end{pmatrix}
\]

with \( Q_1 \) having order \( p \), \( L_1 \) is a unit lower triangular matrix of order \( q \leq p \), and \( D_1 \) is a block diagonal matrix of order \( q \) with diagonal blocks of order 1 or 2. We refer to this as a partial factorization, but allow the case of a complete factorization \( (p = n) \), in which case \( q = n \).

Once the factorization is available, it may be used to compute the partial solutions:

\[
Lx = b, \quad \begin{pmatrix}
D_1 & 0 \\
0 & I
\end{pmatrix} x = b, \quad \text{and} \quad L^T x = b,
\]

and the corresponding equations for \( n \times nrhs \) matrices \( X \) and \( B \).

LAPACK does not offer any routines for partial factorizations but does provide two subroutines for the complete factorization of symmetric indefinite systems. Each of these has potentially serious drawbacks. For \( A \) in packed form, \_sptrf is available. This limits storage requirements but is slow since it does not use blocking. For \( A \) in full form, wasting about half the storage, there is \_sytrf. Both routines employ the Bunch-Kaufmann partial pivoting strategy, which has satisfactory backward stability but the entries of \( L \) may be large, making it unsuitable in some applications (see, for example, [1] and the references therein).

Our aim is to develop a robust and efficient partial (and complete) factorization kernel that employs a stable threshold partial pivoting strategy, limits storage to approximately \( \frac{1}{2}n^2 \), and employs blocking techniques that permit the use of high-level BLAS. Achieving all of these requirements presents a substantial challenge; how we have addressed this is the subject of this talk.
A key part of the talk is to explain the blocking strategy used within our factorization algorithm. Two block approaches to the factorization are considered: a right-left looking algorithm and a recursive algorithm, both employing a partial pivoting strategy that favours the selection of $2 \times 2$ pivots and bounds the entries of $L$. The recursive algorithm is attractive as it has the potential for using level 3 BLAS for very large submatrices and avoiding detailed coding of addressing within blocks. Both approaches have been implemented; numerical results that support our rejection of the recursive algorithm are presented.

The package that we have developed for the partial (or complete) factorization and subsequent solution of symmetric indefinite (possibly singular) systems is HSL\texttt{MA64}. Performance results for HSL\texttt{MA64} on a single core and on eight cores of a multicore machine are given and are shown to compare favourably with the peak performance of _gemm_. For the case $p = n$, results comparing HSL\texttt{MA64} with the LAPACK routine _sytrf_ are also presented. HSL\texttt{MA64} is shown to about 10 per cent faster whilst having the advantages of using about half as much memory (a serious consideration when the multifrontal method is used for solving very large linear systems) and of computing a factorization with $\|L\|$ bounded.

HSL\texttt{MA64} has been employed within our latest multifrontal code and results that demonstrate its effectiveness when used to solve problems from a range of application areas are reported on. Further details may be found in [2].

References


Calibrating Robotic Vision Systems

*Mili Shah*

**Abstract**

With the advent of newer and more technologically advanced robotic vision systems, there is greater need for novel mathematical techniques in order to calibrate these systems. For example, consider cars going down an assembly line. Most of the time the assembly line has to stop in order for the robots to operate on the cars. But if the robots could visually track the cars, then the robots could operate on the car while the assembly line is in motion and thus increase the efficiency of the assembly line. In order for the robots to track the car, each is equipped with cameras that incorporate a robotic vision system. The goal of this proposal is evaluate the accuracy of a specific robotic vision system by comparing the data it gathers from its vision system with data that are collected from a precise sensor system considered ground truth. The problem with comparing these two data streams is that they are not necessarily in the same coordinate system. Therefore, a transform from the robotic vision system’s data stream to the coordinate system of ground truth is necessary. Once this transform is obtained, a metric can be calculated to track how well the underlying robotic vision system works. By ranking these metrics, the optimal system can be chosen. Ultimately, this optimal system will be used as the robotic vision system that allows robots to track.

I am interested in calibrating the six degrees of freedom (6DoF) for an object which track not only the position but also the orientation of the object. These six degrees of freedom represent translations along three perpendicular axes: left and right (along the x axis), forward and backward (along the y axis), and up and down (along the z axis); along with the rotations about those three perpendicular axes (pitch $r_x$, roll $r_y$, and yaw $r_z$). My goal is to find a rotation and translation that best transform one stream of 6DoF data into the another. In order to find such a transformation, I use a matrix representation for the 6DoF data. If the 6DoF representation of an object is represented as $(x, y, z, r_x, r_y, r_z)$ then it may be arranged as a homogeneous matrix

$$H = \begin{pmatrix} R & t \\ 0 & 1 \end{pmatrix},$$

where $t = (x, y, z)^T$ represents the position of the given object and

$$R = R_x R_y R_z$$

with

$$R_x = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(r_z) - \sin(r_z) & 0 \\ 0 & \sin(r_z) & \cos(r_z) \end{pmatrix}, \quad R_y = \begin{pmatrix} \cos(r_y) & 0 & \sin(r_y) \\ 0 & 1 & 0 \\ -\sin(r_y) & 0 & \cos(r_y) \end{pmatrix}, \quad R_z = \begin{pmatrix} \cos(r_z) & -\sin(r_z) & 0 \\ \sin(r_z) & \cos(r_z) & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

represents the orientation of a given object. Given two streams of such 6DoF data,

$$X = \begin{bmatrix} (R_0 t_0) \\ (R_1 t_1) \\ \vdots \\ (R_{n-1} t_{n-1}) \end{bmatrix}, \quad \hat{X} = \begin{bmatrix} (\hat{R}_0 \hat{t}_0) \\ (\hat{R}_1 \hat{t}_1) \\ \vdots \\ (\hat{R}_{n-1} \hat{t}_{n-1}) \end{bmatrix},$$

I will construct the best rotation $R$ and translation $t$ that fits the data. In other words, the best homogeneous matrix $H = (R \ t)$ that minimizes

$$\min_H \|HX - \hat{X}\|^2$$

is constructed. The solution $H$ to the optimization problem (1) involves a two step process:
1. Find the rotation $\mathbf{R}$ that minimizes

$$\min_{\mathbf{R}} \left\| \mathbf{R}(\mathbf{R}_0 \mathbf{t}_0 \ldots \mathbf{R}_{n-1} \mathbf{t}_{n-1}) - (\hat{\mathbf{R}}_0 \hat{\mathbf{t}}_0 \ldots \hat{\mathbf{R}}_{n-1} \hat{\mathbf{t}}_{n-1}) \right\|^2$$

where

$$\mathbf{t}_i = t_i - t \quad \text{and} \quad t = \frac{1}{n} \sum_{i=0}^{n-1} t_i$$

$$\hat{\mathbf{t}}_i = \hat{t}_i - \hat{t} \quad \text{and} \quad \hat{t} = \frac{1}{n} \sum_{i=0}^{n-1} \hat{t}_i$$

2. Set the best transformation

$$\mathbf{t} = \hat{\mathbf{t}} - \mathbf{R} \mathbf{t}$$

where $\mathbf{R}$ is calculated from Step 1.

I will discuss the actual formulation of the rotation $\mathbf{R}$ and translation $\mathbf{t}$. Moreover, I will mathematically prove that the generated $\mathbf{R}$ and $\mathbf{t}$ are indeed the optimal choices that solve the optimization problem (1). In addition, I will present applications of this problem for robotic vision systems.
Convergence Theory for a Restarted GMRES Method with Approximate Deflation Preconditioning

Josef Sifuentes, and Mark Embree

Abstract

The GMRES iterative method approximates a solution to the equation $Ax = b$ via an orthogonal projection onto a Krylov subspace that grows in dimension with each iteration. The periodic restarting of the GMRES process alleviates computational costs and storage requirements, but often causes convergence to stagnate. Accelerating convergence of a restarted GMRES method is widely considered in the literature, see, e.g., [1, 2, 3]. One class of algorithms that aims to do this is deflated preconditioning, which uses knowledge (or estimates) of eigenvectors in order to remove certain components from the initial residual $r_0 = b - Ax_0$. Ehrel et al. and Baglama et al. both show that if one had an orthogonal basis for an $r$-dimensional invariant subspace, then one could construct a preconditioner $M^{-1}$ such that the corresponding $r$ eigenvalues of $M^{-1}A$ are sent to one [1, 2]. They both go on to further analyze convergence of exactly deflated systems.

However, in practice, one does not typically have access to an invariant subspace basis. One does have access to an orthogonal basis for the $r$-dimensional Krylov subspace, generated automatically by the Arnoldi process driving GMRES. In this talk, we show that when one constructs a preconditioner $M^{-1}$ using Arnoldi vectors, then $r - 1$ eigenvalues of $M^{-1}A$ are sent to one. This result is quite remarkable in that it is not a function of how well the Krylov subspace approximates an invariant subspace. However, the magnitude in which the other eigenvalues are perturbed is a function of how well an invariant subspace is approximated.

This begs the question - exactly which invariant subspace is approximated by a Krylov subspace? Stewart’s [5] work on perturbation theory for invariant subspaces provides the analytical framework to understand the answer to that question. We apply Stewart’s analysis to show that if the columns of $V$ are Arnoldi vectors and the residual

$$\rho := \min_{L \in C^{r \times r}} \|AV - VL\|$$

is sufficiently small, then the approximated invariant subspace can be written as a perturbation of the approximate subspace. Furthermore, we show that the approximately deflated system can be written as a perturbation to an exactly deflated system with perturbation norm proportional to the residual $\rho$.

Since convergence of GMRES on exactly deflated systems is described well in [1, 2], we can describe convergence of GMRES on approximately deflated systems by applying results on the stability of GMRES. This talk will apply the work in [4] to show that the residual produced by GMRES applied to the approximately deflated system lags behind that produced by the exactly deflated system by no more than $O(\rho)$.

We demonstrate the effectiveness of approximate deflation preconditioning through an application from penetrable scattering theory. Here the coefficient matrix is a discretization of a shifted compact operator. While the eigenvalues accumulate near 1, several pass close to the origin, causing slow GMRES convergence. Approximate deflation preconditioning effectively deflates the influence of those small eigenvalues, thus accelerating the convergence of a restarted GMRES method.
References


The Schur Complement Method for Nearly-Hermitian Linear Systems: An Effective Solver and Preconditioner

Mark Embree, Josef Sifuentes, Kirk M. Soodhalter, and Daniel B. Szyld

Abstract

We discuss an approach for solving a nearly-Hermitian linear system $Ax = b$, where $A \in \mathbb{C}^{n \times n}$ has the form

$$A = H + \frac{1}{2}FG^* \text{ with } F, G \in \mathbb{C}^{n \times s} \text{ and } s \ll n,$$

where

$$H = \frac{1}{2}(A + A^*) \text{ and } FG^* = A - A^*.$$  

Such systems arise in discretizations of certain integral equations, e.g., wave scattering applications as well as discretizations of self-adjoint differential operators with particular Neumann boundary conditions. Matrices of this form can be interpreted as low-rank modifications of Hermitian matrices. Since we can rewrite $G = 2FC^*$ for an appropriate $C \in \mathbb{C}^{s \times s}$, $A = H + FCF^*$ can be seen as the Schur complement of the $(n + s) \times (n + s)$ matrix,

$$\begin{bmatrix}
H & F \\
F^* & -C^{-1}
\end{bmatrix},$$

and the true solution $x$ satisfies the linear system

$$\begin{bmatrix}
H & F \\
F^* & -C^{-1}
\end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}.$$  

We can thus solve the original system by solving $s + 1$ Hermitian systems of order $n$, e.g., using MINRES, which leads to significant storage savings. We then directly solve one non-Hermitian system to compute $y$. This approach is an application of the Sherman-Morrison-Woodbury identity.

We present numerical results demonstrating our Schur Complement method’s competitiveness with both a standard Krylov method (GMRES) and a short-term recurrence method (IDR). We present theory on stopping criteria and bounds on computed quantities.

We were motivated to think about this problem after discovering that Progressive GMRES [1], which was designed for problems with the structure of (1), was unstable.

Observe that if we have a matrix $A = H + K$ which is a perturbation of a nearly-Hermitian linear system, i.e., $K = \frac{1}{2}FG^* + E$, $\|E\| < \varepsilon$ for $\varepsilon \ll 1$, and $FG^*$ is rank $s$, then the Schur Complement method can be used to approximate the action of $M^{-1}$, $M = H + \frac{1}{2}FG^*$, as a preconditioner for such systems whose skew-Hermitian part is well-approximated by one of low-rank.

These approaches only work when we possess explicit knowledge about the skew-Hermitian part of $A$. In many cases, we may know only that $A$ is nearly-Hermitian or a perturbation thereof. In such circumstances, we can adapt the Schur Complement method to fit into a flexible Krylov method (in this case GMRES). At the $j$th step of a Krylov subspace method, we have built a $j$-dimensional subspace. At the same time, we have built a an operator, $H_j$, which is the restriction and projection of the operator $A$ to the subspace. Since $j$ is small, we have direct access to the skew-Hermitian part of the matrix $H_j$, and we can directly compute some decomposition $H_j - H_j^* = \tilde{F}_j\tilde{G}_j$. We can lift $\tilde{F}_j$ and $\tilde{G}_j$ to the full space producing $F_j$ and $G_j$, approximations to $F$ and $G$, respectively. In the case of our implementation, we use the singular value decomposition $\tilde{U}_j\tilde{\Sigma}_j\tilde{V}_j^* = H_j - H_j^*$,
assigning $\tilde{F}_j = \tilde{U}_j \tilde{\Sigma}_j$ and $\tilde{G}_j = \tilde{V}_j$. Using our Schur complement method, we can approximate the action of $M_j^{-1}$, where $M_j = H + F_j G_j^*$, as a preconditioner at the $j$th step. In practice, we run a fixed number $k$ of GMRES steps before using this strategy in order to build a large enough Krylov subspace to get reasonable approximations to $F$ and $G$. At each subsequent step we compute new approximations $F_j$ and $G_j$ meaning the preconditioner $M_j$ changes at each step. Thus, we have a flexible preconditioning strategy. We present numerical results on systems of dimension $10^6$ demonstrating both the effectiveness and robustness of this strategy.

References

Model reduction seeks to replace a large-scale system of differential or difference equations by a system of substantially lower dimension that has nearly the same response characteristics. There have been a number of very effective approaches developed for model reduction of linear time invariant systems. However, while it is very elegant, little of the linear theory can be applied or readily extended to problems involving model reduction of nonlinear systems. In fact, the linear theory does not even apply to linear time varying systems.

Proper Orthogonal Decomposition (POD) is a well known linear methodology that does apply to nonlinear systems. This approach derives a reduced basis empirically by truncating the Singular Value Decomposition (SVD) of a matrix formed from samples (or “snapshots”) of a typical trajectory. This reduced basis is then used to accomplish dimension reduction by approximating the state variables with an expansion in the reduced basis and then deriving a weak form of the nonlinear system via Galerkin projection with respect to this reduced basis set. In many applications, this approach can be very effective. If the system of nonlinear ordinary differential equations arises from the spatial discretization of a time-dependent parabolic PDE, the number of variables are often reduced by several orders of magnitude (e.g. 15,000 variables reduced to 40).

Unfortunately, this very effective dimension reduction often does not yield a corresponding reduction in actual computation time which is, of course, the ultimate goal of model reduction. The reason for this is that while the number of variables has been greatly reduced, the computational complexity remains that of the original full dimensional system. This is due to the fact that the reduced variables must be used as expansion coefficients in the reduced basis to obtain an approximate state variable (of full size) and the nonlinear right hand side of the must be evaluated at this state variable and then projected orthogonally onto the reduced basis. As a result, the reduced system can be just as time consuming to integrate as the original full dimensional system.

A solution to this complexity problem that will provide a significant complexity reduction comparable to the reduction in variables is available through additional linear algebraic analysis and techniques. The key is to replace the orthogonal projection of POD with an interpolation projection. The result is a methodology called Discrete Empirical Interpolation (DEIM). The approach is a discrete variant of Empirical Interpolation [1] that was originally developed within a Finite Element setting.

DEIM is a modification of POD that reduces complexity of the nonlinear term of the reduced model to a cost proportional to the number of reduced variables obtained by POD. The complexity reduction is a result of the fact that the interpolation projection will only require evaluation of the nonlinear term at a few selected components (The number is proportional to the number of reduced variables in POD). The method applies to arbitrary systems of nonlinear ODEs, not just those arising from discretization of PDEs.

In this talk, the DEIM method will be developed along with a discussion of its approximation properties. Applications from Neural Modeling, Chemically Reacting Flow, Shape Optimization and Porous Media Flow will be presented to illustrate the effectiveness and wide applicability of the DEIM approach. In several of these, the number of variables as well as the computation times are reduced by a factor of 1000.
References

A New Algorithm for the Computation of the Real Distance to Instability

Alastair Spence, and Muzhi Yang

Abstract

A matrix $A$ is said to be stable if all its eigenvalues lie in the left half-plane; otherwise it is unstable. It is a classic problem in numerical linear algebra with many applications, especially in control theory, to find the “distance to instability”, or the “stability radius”, defined by

$$
\min\{\|E\| \mid A + E \text{ is unstable}\}.
$$

An elegant approach to this problem was given by Byers in 1988 (SIAM J. Sci Comput. Vol 9) and his algorithm was later refined by others, for example, Boyd&Balakrishnan 1990 (Systems Control Lett. Vol 14). Using the original Byers formulation, Freitag&Spence (LAA, submitted) developed a new approach to the problem based on computing a 2-dimensional Jordan block in the Byers matrix, using ideas from numerical bifurcation theory.

Even if $A$ is real the perturbing matrix $E$ is usually complex. However, in many physical applications it is of interest to find the smallest real perturbing matrix, whose norm is called the “real stability radius”. A formula to find this value was given in a famous paper by Qiu, Bernhardson et.al. in 1995 (Automatica, Vol 31), though the resulting expression doesn’t lead to an efficient algorithm for its computation.

This talk is about work done by the authors who have extended the ideas in Freitag&Spence and have applied techniques of numerical bifurcation theory to derive a new efficient algorithm to compute the real stability radius of a real stable $A$. 
Large-scale Nonnegative Least-Squares

D. Kim, S. Sra, and I. S. Dhillon

Abstract

We study the fundamental problem of nonnegative least squares, which was introduced by Lawson and Hanson [11] under the name NNLS. Formally, NNLS seeks to solve

$$\min_x f(x) = \frac{1}{2} \|Ax - b\|^2, \quad \text{s.t.} \quad x \geq 0,$$

where $A \in \mathbb{R}^{m \times n}$ is a design matrix (discretized integral operator, measurement matrix, etc.) and $b \in \mathbb{R}^m$ a vector of observations. NNLS arises when apart from a traditional least-squares model, the variables must additionally fulfill nonnegativity constraints [5]. These constraints stem naturally from physical grounds, e.g., when the variables $x$ encode intrinsically nonnegative quantities such as frequency counts (data mining [10]), chemical concentrations (chemometrics [6]), and image intensities or photon counts (astronomy [13, 9]; medical imaging [14]).

Despite its apparent simplicity and close connection to ordinary least-squares, NNLS can be challenging to solve, especially when dealing with large-scale problems. For such problems iterative instead of direct methods are preferable. But the nonnegativity constraints confound out-of-the-box invocations of methods such as conjugate gradients, so more careful algorithms must be developed. Indeed, a variety of methods can be applied to NNLS. Several of these are summarized in [5, Chapter 5]. The main approaches can be roughly divided into three categories: (i) least-distance methods; (ii) active-set approaches; and (iii) general iterative methods.

The first category includes methods developed for linear inequality constrained least-squares, which can be transformed into least distance problems [5]. However, for NNLS such transformations prove to be needlessly expensive. The second category uses active-sets to tackle usually one constraint per iteration. The famous nnls algorithm of Lawson and Hanson [11] is an active set method, and it has been the de facto method for solving (1). Bro and Jong [6] modified nnls to develop Fast-NNLS (fnnls), a method that is often faster than the original Lawson-Hanson algorithm. But fnnls depends on constructing $A^TA$, which makes the method prohibitive for large-scale problems. The third category includes optimization algorithms such as gradient-projection [15, 3], or more powerful approaches such as LBFGS-B [7], TRON [12], and interior point methods [2].

We present an efficient new algorithm that belongs to the third category above. Our algorithm is built on the unconstrained quadratic optimization algorithm of Barzilai and Borwein [1] (hereafter BB), to handle nonnegativity constraints. The resulting algorithm is simple yet efficient. It differs from other constrained BB variants (e.g., [4, 8]) in two key aspects: (i) it uses a specific subset of variables for computing BB steps; and (ii) it scales these steps adaptively to ensure convergence. These differences have strong empirical impact; compared against established convex solvers and specialized NNLS methods, our method exhibits superior performance.

We have extensively experimented with our method, and from these experiments we extract below in Table 1 a sample comparison against fnnls [6], LBFGS-B [7], and SPG [4]. We ran all experiments on a Linux machine, with an Intel Xeon 2.0GHz CPU and 16GB memory. We note that the underlying implementations of LBFGS-B and SPG are in FORTRAN, while fnnls and our method have purely MATLAB implementations.

Ongoing and future work. Encouraged by the promising performance of our method, we are currently extending it to incorporate: (i) pre-conditioning; (ii) multi-core and GPU implementa-
Table 1: Our method compared against others on a matrix of size 9600 × 6400. The true number of “active” constraints (i.e., number of zeros in the optimal solution $x^*$) was 4378. Convergence criterion used by all methods (except FNNLS) was $\|\nabla f_+\|_\infty \leq 10^{-6}$, where the projected-gradient is given by $[\nabla f_+(x)]_i = \min\{0, \partial_i f(x)\}$ if $x_i = 0$, and $[\nabla f_+(x)]_i = \partial_i f(x)$, if $x_i > 0$; note that $\|\nabla f_+(x^*)\|_\infty = 0$ holds at optimality.

<table>
<thead>
<tr>
<th>Method</th>
<th>Time (s)</th>
<th>$f(x)$</th>
<th>#active</th>
<th>$|\nabla f_+|_\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>FNNLS</td>
<td>15,526</td>
<td>8.72e+10</td>
<td>4,738</td>
<td>-</td>
</tr>
<tr>
<td>LBFGS-B</td>
<td>1,516.4</td>
<td>8.72e+10</td>
<td>1,930</td>
<td>5.04 (fails)</td>
</tr>
<tr>
<td>SPG</td>
<td>3,936.5</td>
<td>8.72e+10</td>
<td>2084</td>
<td>16.19 (fails)</td>
</tr>
<tr>
<td>Our Method</td>
<td>290.73</td>
<td>8.72e+10</td>
<td>4,737</td>
<td>9.60e-07</td>
</tr>
</tbody>
</table>

The references include:

When Is Twice Enough?
(The Oblique Case)

G. W. Stewart

Abstract

Let $P$ be a projector (a.k.a. an idempotent matrix). The vector $x = Pa$ is called the projection of $a$ onto the column space $\mathcal{X}$ of $P$, and the vector

$$z = (I - P)a = a - x$$

is called the complementary projection onto the null space $\mathcal{Z}$ of $P$. In most applications it is inefficient to form and apply the complementary projector $I - P$, and we must compute $z$ in the form $z = a - x$. This talk is concerned with the numerically consequences of this complementation procedure.

When $\|P\| = 1$ in the spectral norm, $\mathcal{X}$ and $\mathcal{Z}$ are orthogonal to one another, and the complementation process produces a vector $z$ that is orthogonal to $\mathcal{X}$; i.e., the procedure orthogonalize $a$ with respect to $\mathcal{X}$. If $P$ is represented in the form $XX^*$, where $X$ is an orthonormal basis for $\mathcal{X}$, then the orthogonalization procedure $z = a - X(X^*z)$ is none other than the classical Gram-Schmidt procedure, which has been extensively studied (for a good entry into the literature see [2]).

An undesirable feature of the Gram–Schmidt method is that in the presence of rounding error the vector $z$ can be far from orthogonal to $\mathcal{X}$. This loss of orthogonality is signaled by cancellation in the formation of $z - x$, which magnifies previous rounding errors, which in turn will generally contain components in $\mathcal{X}$. A cure for this problem is to repeat the the procedure on the vector $z$. It has been observed that one reorthogonalization is usually sufficient to produce a vector that is orthogonal to working accuracy — i.e., twice is enough (see [3, p. 110]). But when $a$ is very near $\mathcal{X}$, the process may fail. The cure is to restart with a small vector $\hat{z}$ for which twice is indeed enough. This solution was first described in [1], where a deterministic algorithm for generating $\hat{z}$ was given. In [4] it was shown that a cheaper random algorithm is just as effective.

In this talk, I will show how some of these ideas generalize to oblique projectors, for which $\|P\| > 1$ [5]. Unlike orthogonalization, our ability to achieve nearly exact complementation is limited by the size of $\|P\|$. Specifically, if we take a vector $z \in \mathcal{Z}$ and perturb it slightly to get $\tilde{z} = z + e$, then $P\tilde{z} = Pz + Pe = Pe$; and if $\|P\|$ is large, then so may be $Pe$. Moreover, if $x$ is large, then $x + z$ may not exactly reproduce $a$. But granting these limitation, we will show that two complementations usually give vectors $x$ and $z$ lying in $\mathcal{X}$ and $\mathcal{Z}$ to working accuracy and whose sum reproduces $a$ as well as possible. The analysis also gives conditions under which two complementations can fail. The problem of what to do in that case is an open question.

References


Preconditioning Time-Dependent Optimal Control Problems

Martin Stoll, and Andy Wathen

Abstract

The solution of optimal control problems with partial differential equation (PDE) constraints plays an important role in many applications areas from engineering or the sciences. Using a discretize-then-optimize approach we use finite elements to discretize the objective function and the corresponding PDE and by using a standard Lagrangian approach the first order conditions lead to a saddle point system representing the PDE, the adjoint PDE and the gradient equation. Typically, it will not be feasible to use any direct solver for the saddle point system because of the fine mesh and the 3-dimensional nature of the underlying application.

Many processes in real-world applications can be modeled using time-dependent PDEs such as the heat equation or the time-dependent Stokes equation. We now consider a time-dependent tracking-type functional

\[ J(y,u) = \frac{1}{2} \int_0^T \int_{\Omega_1} (y - \bar{y})^2 \, dx \, dt + \frac{\beta}{2} \int_0^T \int_{\Omega} (u)^2 \, dx \, dt + \frac{\gamma}{2} \int_{\Omega_1} (y(T) - \bar{y}(T))^2 \, dx \]  

or variations thereof, where \( \Omega_1 \subseteq \Omega \) are bounded domains in \( \mathbb{R}^d \) with \( d = 2, 3 \) and \( \bar{y} \) is a desired state. Additionally, for the state \( y \) and the control \( u \) state equations such as the heat equation

\[ y_t - \triangle y = u \quad \text{in} \quad [0,T] \times \Omega \]  
\[ y(t) = 0 \quad \text{in} \quad \partial \Omega \]  
\[ y(0,.) = y_0 \quad \text{in} \quad \Omega \]  

or the time-dependent Stokes equation

\[ y_t - \nu \triangle y + \nabla p = u \quad \text{in} \quad [0,T] \times \Omega \]  
\[ -\nabla \cdot y = 0 \quad \text{in} \quad [0,T] \times \Omega \]  
\[ y(t,\cdot) = g(t) \quad \text{on} \quad \partial \Omega, t \in [0,T] \]  
\[ y(0,\cdot) = y_0 \quad \text{in} \quad \Omega \]  

have to be satisfied. Here, we will focus on the so-called one-shot approach where the iterative solver computes for all time-steps at once. After discretization the appropriate one-shot form for the forward heat equation is now given by

\[
\begin{bmatrix}
L & 0 & 0 & 0 & 0 \\
-M & L & 0 & 0 & 0 \\
0 & -M & L & 0 & 0 \\
0 & 0 & \ddots & \ddots & 0 \\
0 & 0 & 0 & -M & L
\end{bmatrix}
\begin{bmatrix}
y^1 \\
\vdots \\
y^N
\end{bmatrix} - \mathcal{M}u =
\begin{bmatrix}
My^0 \\
0 \\
0 \\
\vdots \\
0
\end{bmatrix}
\]

\[ \Leftrightarrow Ky - \mathcal{M}u = d, \]

222
where $L = M + \tau K$ with $M$ the mass and $K$ the stiffness matrix. Now a standard Lagrangian approach will lead to a saddle point problem of vast dimensions as we have the spatial times the temporal discretization determining the matrix size, e.g.

$$
\begin{bmatrix}
\tau \mathcal{M} & 0 & -\mathcal{K}^T \\
0 & \beta \tau \mathcal{M} & \tau \mathcal{M} \\
-\mathcal{K} & \tau \mathcal{M} & 0
\end{bmatrix}
$$

(12)

In our talk, we will investigate the efficient solution of the above saddle point problem by using a pre-conditioned version of MINRES. We will propose suitable approximations to the Schur-complement of the saddle point matrix that will not require the solution with the discretized PDE operator. We will also present an eigenvalue analysis that guarantees the independence of the preconditioned system with respect to the mesh parameter $h$. In addition to our theoretical work, we will present a set of numerical examples that illustrate the competitiveness of our approach.
Matching Moments and Matrix Computations

Jörg Liesen, Zdeněk Strakoš and Petr Tichý

Abstract

Matching moments is inherently linked with numerical quadrature, continued fractions and orthogonal polynomials. Its relationship to matrix computations has a remarkable history. In this talk we will discuss how the classical topics of the 19th century mathematics found so widespread applications in solving modern computational problems formulated via matrices, and show that awareness of the underlined relationships can be beneficial in constructing and analysis of modern methods and algorithms.

The concept of moments arose with the work of Chebyshev, Markov and Stieltjes in the second half of the 19th century. The related numerical quadrature is even older. It started with Gauss (see the 1814 paper [1]), with further founding contributions due to Jacobi, Christoffel, Markov, Stieltjes and many others. The related fundamental concept of continued fractions can essentially be rooted back to Euclid and other ancient mathematicians. It is present in many mathematical disciplines such as number theory, approximation theory, probability and statistics, numerical quadrature, spectral theory, and modern matrix computations; see, e.g., [2]. The same is true for a slightly more recent concept of orthogonal polynomials. Among very many mathematicians who contributed in a fundamental way to the related developments one should not miss Brouncker and Wallis giving in 1655 what we now call the Stieltjes three-term recurrences, Euler with his overwhelming 18th century work including infinite series expansions of continued fractions, and Stieltjes with his analytic theory of continued fractions described in a truly fascinating paper [3] published in 1894. That paper also presented, as a by-product, a (Riemann-Stieljes) generalization of the Riemann integral, and a formal description of the moment problem together with its complete solution.

We can observe an impact of the related results in forming foundations of functional analysis by Hilbert in 1906 - 1912, as well as in forming mathematical foundations of quantum mechanics by von Neumann in 1927 - 1932. Continued fractions and Jacobi matrices are still important in spectral theory of operators in mathematical physics and they are used in physics and computational chemistry.

The original work of Krylov from 1931 focused on computation of the minimal polynomial via a transformation of a secular equation [4], and it refers in a substantial way to the work of Jacobi [5] from 1846. Its algebraic formulation, with using what we now call the Krylov sequence, was given by Gantmacher in 1934 [6]. In modern computational mathematics, sciences and engineering, Krylov subspace methods and matching moments model reduction (in approximation of large scale dynamical systems and elsewhere) can be viewed as nothing but a translation of the classical concepts mentioned above to the language of large scale matrix computations. Surprisingly, several important works which made these links transparent remained almost unknown; see, in particular, the work of Vorobyev [7] and the description in [8]. More details are given in the review [9].

One can take the links mentioned above simply as a material worth of historical essays. We will, however, argue that there is more to learn here. Looking back has a positive impact on modern computations; examples can be found, e.g., in [10, 11]. Viewing relevant matrix computations as matching moments is inspirational and useful for understanding of behavior of methods and algorithms, and sometimes it can be useful for examining validity of approaches used in modern matrix computations. We will demonstrate this on several examples partially presented in the recent works [12, 13, 14]. In particular, we will address the questions of cost evaluation in Krylov
subspace iterations and of efficient numerical approximation of the bilinear form $c^* A^{-1} b$. Besides the analytic aspects we will show importance of numerical stability considerations.

References


A Divide and Conquer Method for the SVD of a Banded Matrix

Xiaobai Sun

Abstract

A divide and conquer method is introduced for computing the singular-value system (singular values and singular vectors) of a banded matrix. A banded triangular matrix may arise, for example, as an intermediate form in the singular-value decomposition (SVD) process when the successive band reduction approach is used in the reduction to the bidiagonal form for increasing data locality and computation concurrency. Instead of taking the final band reduction to the bidiagonal form as a whole (which may or may not be decoupled numerically), this new method splits, recursively, the banded matrix into the sum of a decoupled banded matrix and a lower-rank matrix. It then obtains the singular-value system by coupling the decoupled and smaller singular-value systems with a low-rank update.

This SVD method for a matrix of size $N$ with bandwidth $b$ takes $O(bN^2 \log(N))$ arithmetic operations, faster than other existing SVD methods for a banded matrix. The singular vector matrices can be represented in compressed form and take only $O(bN \log(N))$ floating point numbers in memory space. For very large $N$, when element-wise approximation by the fast multipole method, for example, is used in the coupling process, the arithmetic complexity can be further reduced to $O(bN \log^2(N))$. The method leads straightforwardly to parallel computation, not limited to the parallel fashion in fine-grain pipeline as in the reduction of a banded matrix to the bidiagonal form.

The introduced method draws upon the advanced theory and algorithms for the symmetric tridiagonal eigenproblem, especially, since the work by Cuppen in 1981, three decades ago.
Stable Computation of the CS Decomposition: Simultaneous Bidiagonalization

Brian D. Sutton

Abstract

At Householder Symposium XVII, we presented an algorithm for the complete 2-by-2 CS decomposition (CSD). This is the original decomposition of Stewart’s 1977 paper, as opposed to the 2-by-1 form that underlies the generalized singular value decomposition. In 2008, numerical stability was suggested but not proved. In this talk, we present a proof.

The CSD exists for any unitary matrix partitioned into a 2-by-2 block structure. It simultaneously diagonalizes the four blocks with a pair of block-diagonal unitary transformations, and thus each unitary transformation serves as a left or right singular vector matrix for two different blocks at the same time. This extensive sharing of singular vectors explains the difficulty of computation. Great care must be taken to choose singular vectors consistently over all four blocks, even when the singular vectors are ill-determined by any one block individually.

The new CSD algorithm proceeds in two phases. In the first phase, the four blocks are simultaneously bidiagonalized, and in the second phase, they are iteratively diagonalized through QR iteration. As presented at the previous Householder Symposium, the bidiagonalization procedure uses half as many Householder reflections as one might first guess, taking advantage of orthogonality to work on two rows or columns at the same time. The question remaining in 2008 was, What happens when the input matrix is not exactly unitary? In practice, certain entries are small rather than zero. Are they small enough to neglect?

In this talk, we answer yes, but our argument takes us back to the beginning. We develop a modified algorithm that enables a much simplified stability analysis. Surprisingly, the modified algorithm violates the matrix partitioning initially, but these operations are later inverted on paper. In code, the modified algorithm is numerically equivalent to the original algorithm, suffering from exactly the same roundoff errors. We show that the computed solution is the exact solution to a small perturbation of the input. The norm of the perturbation is proportional to unit roundoff and to the distance from orthogonality of the input matrix.
Petrov Galerkin View of IDR (and BiCGStab)

Valeria Simoncini and Daniel B. Szyld

Abstract

The IDR(s) method of Sonneveld and van Gijzen [SIAM J. Sci. Comput., 31:1035–1062, 2008] is probably today’s most effective short-term recurrence iterative method for the solution of (non-symmetric) linear systems of equations. Like BiCGStab, it is a short-term recurrence method in which the approximation at each step lies in an affine Krylov subspace of growing dimension. In fact, BiCGstab has been shown to be mathematically equivalent to a special case of IDR. IDR was originally designed so as to have the residuals at each step lying in a sequence of shrinking subspaces. Thus its name IDR, which stands for induced dimension reduction.

In this contribution, we describe the IDR class of methods as classical Krylov subspace methods satisfying a Petrov-Galerkin condition using an appropriately chosen left subspace. These left subspaces are rational Krylov spaces. It should be noted that the rational Krylov spaces are built implicitly, without the solution of any additional linear systems (i.e., without an inverse). This is a new interpretation of the IDR(s) method, which helps understand its convergence properties. This analysis then permits us for the first time to also view BiCGStab as a Petrov-Galerkin method.

The connection with rational Krylov spaces inspired us to look a new possible values for the poles of the defining rational functions. We present a new version of IDR, where the poles of the rational function are chosen as certain Ritz values. We call this new version Ritz-IDR, and we present several experiments illustrating its effectiveness.

Other choices of poles may be explored as well.
Preconditioners for PDE-constrained Problems with Nonlinear PDEs in the Constraints

Nick Gould, Dominique Orban, Sue Thorne and Andy Wathen

Abstract

Consider a partial differential equation (PDE) of the general form \( c(y, u) = 0 \) subject to boundary conditions \( \alpha_1 y + \alpha_2 \frac{\partial y}{\partial x} = \mathbf{d} \) over a domain \( \Omega \), where \( \alpha_1, \alpha_2, u \) and \( d \) are given. The iterative solution of the discretized PDE with appropriate preconditioner is now a mature area for many classes of PDEs. However, in many applications, the value of \( u \) and/or \( d \) can be altered and, hence, a problem of great importance is to find \( u \) and/or \( d \) such that \( y \) is close to (or far from) a particular state. For example, in distributed control, a control \( u \) is computed such that the state \( y \) (satisfying the PDE and boundary conditions) approximates a target state \( \hat{y} \) over a domain \( \hat{\Omega} \subset \Omega \). Mathematically, the problem can be expressed as

\[
\begin{align*}
\min_{y, u} & \quad \frac{1}{2} \|y - \hat{y}\|_{L^2(\hat{\Omega})}^2 + \beta \|u\|_{L^2(\Omega)}^2 \\
\text{subject to} & \quad c(y, u) = 0, (x \in \Omega), \\
& \quad \alpha_1 y + \alpha_2 \frac{\partial y}{\partial x} = \mathbf{d}, (x \in \partial\Omega),
\end{align*}
\]

where we have added a regularization term \( \beta \|u\|_{L^2(\Omega)}^2 \) to the cost functional to ensure that the problem is well-posed and, through careful choice of \( \beta \), we hope to keep \( \|y - \hat{y}\|_{L^2(\hat{\Omega})} \) small and compute a control \( u \) that is physically reproducible.

If we consider the Poisson problem \( c(y, u) := -\nabla^2 y - u \) and discretize (1) (using the same discretization for \( y \) and \( u \)), we obtain

\[
\begin{align*}
\min_{y, u} & \quad \frac{1}{2} (y - \hat{y})^T \mathbf{P} \mathbf{M} \mathbf{P} (y - \hat{y}) + \beta u^T \mathbf{M} u \\
\text{subject to} & \quad \mathbf{K} y - \mathbf{M} u = \mathbf{d},
\end{align*}
\]

where \( K \) and \( M \) are, respectively, the stiffness (discrete Laplacian) and mass matrices associated with our chosen discretization method, and \( P = \text{diag}(p_i) \) with

\[
p_i(x) = \begin{cases} 0 : & x \notin \hat{\Omega}, \\ 1 : & x \in \hat{\Omega}. \end{cases}
\]

The matrices \( M, K \) and \( P \) all have dimension \( N \times N \). If we incorporate the constraints into the cost functional by use of Lagrange multipliers, then we wish to solve the problem

\[
\begin{align*}
\min_{y, u, \lambda} & \quad \frac{1}{2} (y - \hat{y})^T \mathbf{P} \mathbf{M} \mathbf{P} (y - \hat{y}) + \beta u^T \mathbf{M} u + \lambda^T (\mathbf{K} y - \mathbf{M} u - \mathbf{d}),
\end{align*}
\]

where \( \lambda \) is the vector of Lagrange multipliers. The optimality (KKT) conditions for (2) are

\[
\begin{bmatrix}
2\beta M & 0 & -M \\
0 & \mathbf{P} \mathbf{M} \mathbf{P} & \mathbf{K} \\
-M & \mathbf{K} & 0
\end{bmatrix}
\begin{bmatrix}
u \\
y \\
\lambda
\end{bmatrix}
=
\begin{bmatrix}
0 \\
\mathbf{P} \mathbf{M} \hat{u} \\
\mathbf{d}
\end{bmatrix}
\]

and, hence, solving this linear system of equations with dimension \( 3N \times 3N \) is equivalent to solving the discretized version of (1).
Instead, let $c(y, u) := -\nabla \cdot ((1 + y^2) \nabla y) - u$, then (incorporating the constraints into the cost function using Lagrange multipliers) we wish to solve the discretized problem

$$\min_{y, u, \lambda} \frac{1}{2} (y - \hat{y})^T PMP (y - \hat{y}) + \beta u^T Mu + \lambda^T (F(y) - Mu - d),$$

where $F(y) = Ky + L(y)y$ is the discretized version of $-\nabla \cdot ((1 + y^2) \nabla y)$. Note that the KKT conditions will now be nonlinear. If we use Newton’s method or the trust-funnel method to solve this minimization problem, we will need to solve a sequence of systems of the form

$$\begin{bmatrix}
2\beta M & 0 & -M \\
0 & PMP + \sum_{j=1}^N \lambda_j \nabla^2 F_j(y) & J(y)^T \\
-M & J(y) & 0
\end{bmatrix}
\begin{bmatrix}
\delta u \\
\delta y \\
\delta \lambda
\end{bmatrix}
= \begin{bmatrix}
M(\lambda - 2\beta u) \\
M \hat{u} - My - J(y)^T \lambda \\
Mu - F(y)
\end{bmatrix},$$

(4)

where $J(y)$ is the Jacobian of $F(y)$ and

$$F(y) = [ F_1(y)^T, F_2(y)^T, \ldots, F_N(y)^T ]^T.$$

Note that we can easily eliminate $u$ from (3) and $\delta u$ from (4) to obtain the reduced $2N \times 2N$ systems

$$\begin{bmatrix}
PMP & K \\
K & -\frac{1}{2\beta} M
\end{bmatrix}
\begin{bmatrix}
y \\
\lambda
\end{bmatrix}
= \begin{bmatrix}
PMP \hat{u} \\
d
\end{bmatrix}$$

(5)

and

$$\begin{bmatrix}
PMP + \sum_{j=1}^N \lambda_j \nabla^2 F_j(y) & J(y)^T \\
-J(y) & -\frac{1}{2\beta} M
\end{bmatrix}
\begin{bmatrix}
\delta y \\
\delta \lambda
\end{bmatrix}
= \begin{bmatrix}
\frac{M(\lambda - 2\beta u)}{2\beta M \lambda - F(y)} \\
\frac{M \hat{u} - My - J(y)^T \lambda}{2\beta M \lambda - F(y)}
\end{bmatrix}.$$  

(6)

We shall investigate the use of constraint preconditioners with the project preconditioned conjugate gradient method and block-diagonal preconditioners with the SYMMBK method to solve (3)–(6). The preconditioners are efficient to apply ($O(N)$ operations for each solve with the preconditioner) and the rate of convergence of the iterative linear solver is independent of the mesh size of our discretization. For our proposed constraint preconditioners, the theoretical rate of convergence is the same for the original system and the reduced system but each iteration of the PPCG method requires more operations for the reduced system. If $\hat{\Omega} \subset \Omega$, the eigenvalues of the block-diagonally preconditioned system for the reduced system are more tightly clustered together than those of the original system with the proposed block-diagonal preconditioner. Additionally, as $\hat{\Omega}$ gets smaller, the eigenvalues of the preconditioned systems cluster more for the reduced problem but they spread out for the original system. Hence, we will expect the rate of convergence of SYMMBK to be faster for the reduced system. In addition, each iteration of the SYMMBK method is, on average, cheaper when it is applied to the reduced system. We compliment our theoretical results with a number of numerical examples.
On Best Approximation by Polynomials of Matrices

Vance Faber, Jörg Liesen, Petr Tichý

Abstract

An important part of approximation theory is concerned with the approximation of a given function $f$ on some (compact) set $\Omega$ in the complex plane by polynomials. Classical results in this area deal with the best approximation problem

$$\min_{p \in \mathcal{P}_m} \| f - p \|_\Omega \quad \text{where} \quad \| g \|_\Omega = \max_{z \in \Omega} |g(z)|,$$

and $\mathcal{P}_m$ is the set of polynomials of degree at most $m$.

Scalar approximation problems of the form (1) have been studied since the mid 1850s. Accordingly, numerous results on existence and uniqueness of the solution as well as estimates for the value of (1) are known. Here we consider a problem that at first sight looks similar, but apparently is much less understood: Let $f$ be a function that is analytic in a neighbourhood of the spectrum of a given matrix $A \in \mathbb{C}^{n \times n}$, so that $f(A)$ is well defined, and let $\| \cdot \|$ be a given matrix norm. Consider the matrix approximation problem

$$\min_{p \in \mathcal{P}_m} \| f(A) - p(A) \|.$$

We ask two basic questions: Does this problem have a unique solution? Can we understand some properties of polynomials that solve this kind of best approximation problems?

An answer to the first question depends of course on the norm used in (2). If the norm is known to be strictly convex, as for example the Frobenius norm, then (2) is guaranteed to have a uniquely defined solution as long as the value of (2) is positive. A useful matrix norm that is met in many applications is the matrix 2-norm (spectral norm), which for a given matrix $A$ is equal to the largest singular value of $A$. This norm is not strictly convex, and thus the general result on uniqueness of best approximation in linear spaces with a strictly convex norm does not apply. In this presentation we will consider matrix approximation problems in the matrix 2-norm.

It is well known that when the function $f$ is analytic in an open neighborhood of the spectrum of the matrix $A \in \mathbb{C}^{n \times n}$, then $f(A)$ is a well-defined complex $n \times n$ matrix. In fact, $f(A) = p_f(A)$, where $p_f$ is a polynomial that depends on the values and possibly the derivatives of $f$ on the spectrum of $A$. Therefore, $f$ in (2) can be thought to be a polynomial of degree, say, $m + \ell + 1$ ($m \geq 0$, $\ell \geq 0$). Then, as we show in [4], the problem (2) can equivalently be written in the form

$$\min_{h \in \mathcal{P}_m} \| A^{m+1} g(A) - h(A) \|,$$

where $g$ is a given polynomial of degree at most $\ell$. We can also consider a related problem

$$\min_{g \in \mathcal{P}_\ell} \| A^{m+1} g(A) - h(A) \|,$$

where $h$ is a given polynomial of degree at most $m$, and the best $g \in \mathcal{P}_\ell$ is sought.

Special cases of problems (3) and (4) have been considered by Greenbaum and Trefethen [2] in the context of convergence of Krylov subspace methods for solving linear systems and eigenvalue problems. In particular, they considered the approximation problems

$$\min_{h \in \mathcal{P}_m} \| A^{m+1} - h(A) \| \quad \text{and} \quad \min_{g \in \mathcal{P}_\ell} \| Ag(A) - I \|,$$
the first one called the ideal Arnoldi approximation problem, and the second one called ideal GMRES approximation problem. Greenbaum and Trefethen seem to be the first who studied existence and uniqueness of polynomials that solve the problems (5). In particular, in [2] it is shown that if the minima in (5) are nonzero, and if $A$ is nonsingular in the case of ideal GMRES, then both problems (5) have a unique minimizer. In the first part of this talk we present the results of our paper [4] and generalize results of Greenbaum and Trefethen to problems of the form (3) and (4):

*Provided that the minimum in (3) is nonzero, the problem (3) has a unique minimizer. Provided that the minimum in (4) is nonzero and $A$ is nonsingular, the problem (4) has a unique minimizer.*

Note that the assumption of nonsingularity in the second case is in general necessary.

In a later paper, Toh and Trefethen [5] have called the polynomial that solves the problem

$$\min_{h \in \mathcal{P}_m} \|A^{m+1} - h(A)\| = \min_{p \in \mathcal{M}_{m+1}} \|p(A)\| \equiv \varphi_{m+1}(A)$$

the $(m+1)$st Chebyshev polynomial of $A$. Here $\mathcal{M}_{m+1}$ denotes the class of monic polynomials of degree $m + 1$. The reason for this terminology is the following: When the matrix $A$ is normal, i.e. unitarily diagonalizable, problem (6) for the matrix 2-norm becomes a scalar approximation problem of the form (1), with $\Omega$ being the set of eigenvalues of $A$, and the resulting monic polynomial is the $(m+1)$st Chebyshev polynomial on the (discrete) set of eigenvalues of $A$. Apart from the work of Greenbaum, Toh and Trefethen, and some further numerical examples in the recent book [3], very little has been published on Chebyshev polynomials of matrices, let alone the more general problem (2).

In second part of the talk we will present results of our recent paper [1] that studies general properties of Chebyshev polynomials of matrices. In some cases, these properties turn out to be generalizations of well known properties of Chebyshev polynomials of compact sets in the complex plane. For example, it is well known that Chebyshev polynomials for compact sets are characterized by alternation properties. A similar property can also be shown for Chebyshev polynomials of block diagonal matrices: We show that the minimal norm $\varphi_{m+1}(A)$ is attained on several blocks simultaneously. We also present explicit formulas of the Chebyshev polynomials of certain classes of matrices, including Jordan blocks, perturbed Jordan blocks and special classes of bidiagonal matrices, and explore the relation between Chebyshev polynomials of these classes of matrices and Chebyshev polynomials of sets in the complex plane.

**References**


A Reliable Algorithm for the Complete Solution of Quadratic Eigenvalue Problems

Sven Hammarling, Christopher J. Munro, Françoise Tisseur

Abstract

Eigensolvers for quadratic eigenvalue problems (QEPs) \( Q(\lambda)x = 0, \ y^*Q(\lambda) = 0 \), where

\[
Q(\lambda) = \lambda^2 A_2 + \lambda A_1 + A_0, \quad A_i \in \mathbb{C}^{n \times n}
\]

are often absent from the numerical linear algebra component of software libraries, leaving users to choose a \( 2n \times 2n \) linearization \( L(\lambda) = A - \lambda B \) of \( Q(\lambda) \) then call an eigensolver for generalized eigenproblems and finally recover the eigenvectors of \( Q(\lambda) \) from that of the linearized problem \( L(\lambda)z = 0, \ w^*L(\lambda) = 0 \). However in doing so, it is important to understand the influence of the linearization process on the accuracy and stability of the computed solution. Indeed solving the QEP by applying a backward stable algorithm (e.g., the QZ algorithm) to a linearization can be backward unstable for the QEP [9]. Also, unless the block structure of the linearization is respected (and it is not by standard techniques), the conditioning of the solutions of the larger linear problem can be worse than that for the original quadratic, since the class of admissible perturbations is larger. For example, eigenvalues that are well conditioned for \( Q(\lambda) \) may be ill conditioned for \( L(\lambda) \) [7], [8]. For these reasons, the numerical solution of QEPs requires special attention.

In this work we present a general purpose eigensolver for dense QEPs that incorporates the most recent contributions on the numerical solution of QEPs, namely a scaling of the eigenvalue parameter prior to the computation (Fan, Lin and Van Dooren scaling [3]), and the use of a linearization with favorable conditioning and backward stability properties [5], [6], [7], [8].

Our algorithm is also based on new analysis. The early stages take advantage of the block structure of the chosen linearization, thereby reducing the computational cost. A preprocessing step reveals the zero and infinite eigenvalues contributed by singular leading and trailing matrix coefficients and deflates them. This preprocessing step, based on QR factorizations with column pivoting, is backward stable. The deflated generalized eigenproblem is then solved by the QZ algorithm. We note that if infinite and zero eigenvalues are not extracted before starting the QZ steps then they may never be detected due to the effect of rounding errors in floating point arithmetic.

Backward stability is guaranteed for problems that are not too heavily damped, i.e.,

\[
\|A_1\|_2 \leq \sqrt{\|A_0\|_2\|A_1\|_2}
\]

(1)

(many practical problems have this property—see the NLEVP collection [2]) or for eigenpairs whose eigenvalues have magnitude of order 1.

For heavily damped problems, the algorithm attempts to diagonally scale \( Q(\lambda) \) prior to the eigenvalue scaling so that \( D_1Q(\lambda)D_2 \) is not too heavily damped. The construction of \( D_1 \) and \( D_2 \) is based on Betcke’s weighted scaling [1], which was originally derived for improving the conditioning of eigenvalues of general matrix polynomials.

We also analyze Gaubert and Sharify’s tropical scaling [4] of the eigenvalue parameter, which is derived from the tropical roots of the tropical polynomial

\[
p(\lambda) = \max(\|A_0\|_2, \|A_1\|_2\lambda, \|A_2\|_2\lambda^2)
\]

233
corresponding to $Q(\lambda)$. When (1) holds, $p(\lambda)$ has a double tropical root defining a scaling of $\lambda$ identical to that of Fan, Lin and Van Dooren, which our algorithm uses by default. For heavily damped problems, the tropical roots define two scalings. We show that one of them guarantees backward stability for large eigenvalues and the other one has a similar property but for small eigenvalues. Because backward stability depends on the magnitude of $\lambda$, tropical scaling is left as an option.

Our algorithm is implemented in a MATLAB function \texttt{quadeig}, which makes use of functions from the NAG Toolbox for MATLAB, and in FORTRAN 95 for possible inclusion in LAPACK and the NAG library.

How does our implementation compare with the MATLAB function \texttt{polyeig}? The latter function computes eigenvalues and optionally right eigenvectors and eigenvalue condition numbers of matrix polynomials of arbitrary degree. For quadratics, backward stability is guaranteed only when $\|A_0\|_2 \approx \|A_1\|_2 \approx \|A_2\|_2 \approx 1$. Our MATLAB function \texttt{quadeig} is restricted to quadratic matrix polynomials but can be extended to matrix polynomials of higher degrees. It returns eigenvalues and optionally right/left eigenvectors, eigenvalue condition numbers and backward errors of right and left eigenpairs. Backward stability is guaranteed for a much wider class of quadratics (those that are not too heavily damped). But thanks to appropriate scaling strategies, \texttt{quadeig} is very likely to compute eigenpairs with small backward errors for heavily damped problems. Moreover, \texttt{quadeig} can be much faster than \texttt{polyeig} for problems with singular trailing and or leading coefficient matrices due to our deflation strategy.

References


Projected Residuals and Backward Error Estimates in LSQR

Xiao-Wen Chang, Pavel Jiránek, Chris Paige, and David Titley-Peloquin

Abstract

This presentation is mainly based on material published in [1] and [4]. We consider the iterative solution of large sparse least squares (LS) problems,

\[
\min_x \|b - Ax\|_2,
\]

where \(A \in \mathbb{R}^{m \times n}\) has full column rank and \(b \in \mathbb{R}^m\). Specifically, we focus on the design and implementation of reliable stopping criteria for the widely-used algorithm LSQR of Paige and Saunders [5]. First we review why certain projections of the residual vector are good measures of convergence. Then we show how these projections can be estimated efficiently at every iteration of LSQR.

It is well known that \(x^* \) solves (1) if and only if \(A^T(b - Ax^*) = 0\). Let \(x_k\) denote an approximate LS solution, for example, the \(k\)-th iterate of LSQR. It has often been suggested (see e.g. [1, §2]) that \(x_k\) be considered an acceptable computed solution when, for chosen tolerances \(\alpha \geq 0\) and \(\beta \geq 0\),

\[
\exists \ E, f \quad \text{such that} \quad (A + E)^T[b + f - (A + E)x_k] = 0, \quad \frac{\|E\|_F}{\|A\|_F} \leq \alpha, \quad \frac{\|f\|_2}{\|b\|_2} \leq \beta. \quad (2)
\]

In other words, we should accept \(x_k\) when it is the exact solution of a LS problem within a specified range of normwise relative errors in the data.

Consider the following backward error problem:

\[
\mu_k(\theta) = \min_{E,f} \left\{ \|E,f\|_F : (A + E)^T[b + f - (A + E)x_k] = 0 \right\}. \quad (3)
\]

In the above, \(\theta > 0\) is a scaling parameter that allows for a different weight on the perturbations in \(A\) and \(b\). Define \(\theta = \frac{\alpha \|A\|_F}{\beta \|b\|_2}\). It was shown in [1, §6] that if

\[
\mu_k(\tilde{\theta}) \leq \alpha \|A\|_F \quad (4)
\]

then (2) holds, while if (2) holds then \(\mu_k(\tilde{\theta}) \leq \sqrt{2}\alpha \|A\|_F\). Thus (4) is essentially necessary and sufficient for (2) to hold. The condition (4) was proposed in [1, §6] as a stopping criterion for the iterative solution of LS problems. This requires the computation of \(\mu_k(\theta)\), or of an estimate of \(\mu_k(\theta)\), at every iteration.

Waldén, Karlson, and Sun [7, §2] showed that

\[
\mu_k = \min \left\{ \omega_k, \sigma_{\min} \left( A, \omega_k \left( I - r_k r_k^\dagger \right) \right) \right\}, \quad (5)
\]

where \(\sigma_{\min}(\cdot)\) denotes the smallest singular value, \(\dagger\) denotes the Moore-Penrose generalized inverse, \(r_k \equiv b - Ax_k\), and \(\omega_k \equiv \frac{\theta \|r_k\|_2}{\sqrt{1 + \theta^2 \|x_k\|_2^2}}\). The above expression for \(\mu_k\) involves the smallest singular value of an \(m \times (n + m)\) matrix, which is much too expensive to compute directly in practise. One contribution of the present work is the following. Using the secular equation (see e.g. [3, §2]) we show that if \(\mu_k < \omega_k\), then \(\mu_k\) is given implicitly by

\[
\mu_k = \frac{\omega_k}{\|r_k\|_2} \left\| \begin{bmatrix} A / \sqrt{\omega_k^2 - \mu_k^2 I} \\ A / \sqrt{\omega_k^2 - \mu_k^2 I} \end{bmatrix} \begin{bmatrix} r_k \\ 0 \end{bmatrix} \right\|_2. \quad (6)
\]
From this expression we can easily show that
\[
\frac{\omega_k}{\|r_k\|_2} \left\| \begin{bmatrix} A \\ \omega_k I \end{bmatrix} \begin{bmatrix} A \\ 0 \end{bmatrix}^\dagger \begin{bmatrix} r_k \\ 0 \end{bmatrix} \right\|_2 \leq \mu_k \leq \frac{\omega_k}{\|r_k\|_2} \|AA^\dagger r_k\|_2.
\] (7)

The above lower bound is known in the literature as an asymptotic estimate of \(\mu_k\); see e.g. [2, §2]. The above upper bound appeared in [6, §3.4]. The expression (6) allows us to derive upper bounds on the relative errors between \(\mu_k\) and the projections in (7).

Both bounds in (7) are too expensive to compute to be used directly in practise. Another contribution of this work is the following. We give estimates of \(\mu_k\) analogous to those in (7) that can be computed efficiently at every iteration of the algorithm LSQR. At the \(k\)-th step of LSQR, lower-bidiagonal matrices \(B_k \in \mathbb{R}^{(k+1) \times k}\) and \(\overline{B}_k \in \mathbb{R}^{(k+1) \times (k+1)}\) have (in theory) been produced. We show how \(B_{k+d}\) and \(\overline{B}_{k+d}\) (produced at step \(k + d\) of LSQR) can be used to estimate \(\mu_k\) (obtained at step \(k\) of LSQR). Specifically, we show that

\[
\min \left\{ \omega_k, \sigma_{\min} \left( \begin{bmatrix} B_{k+d}, \omega_k(I - t_k t_k^\dagger) \end{bmatrix} \right) \right\} \leq \mu_k \leq \min \left\{ \omega_k, \sigma_{\min} \left( \begin{bmatrix} \overline{B}_{k+d}, \omega_k(I - t_k t_k^\dagger) \end{bmatrix} \right) \right\},
\]

where \(t_k \in \mathbb{R}^{(k+d+1)}\) is available at step \(k + d\) of LSQR. Each of the above bounds is itself a backward error of the form (5), and for each we obtain estimates analogous to those in (7). The cost of computing these estimates is linear in \(k + d\). Our estimates can therefore be used instead of \(\mu_k\) in (4), making the stopping criterion (4) suitable for use in practise.

References

On the Way Towards Robust Algebraic Preconditioners

Rafael Bru, José Mas, José Marín, and Miroslav Tůma

Abstract

Rapidly increasing sizes of discrete problems arising from solving partial differential equations in three dimensions imply strong need for fast and efficient iterative methods with moderate memory demands. It is well-known that iterative methods must be preconditioned in order to be really useful. Algebraic preconditioners represent a large group of approaches which are important both theoretically and practically, despite their generality and possible lack of specificity for solving particular problems. In addition, such preconditioners are often combined with the problem-based solving strategies, possibly enhanced by the multilevel framework or used to solve non-PDE applications.

In our talk we deal with the problem of robustness of algebraic preconditioners, in particular we consider direct and inverse incomplete decompositions. The development of these decompositions, e.g., standard direct incomplete Cholesky and LU decompositions is traditionally well connected with our community. Their subsequent improvements from early stencil-based relaxed factorizations towards fully general techniques with additional enhancements enabled to successfully solve important applications. The role of both factorized and non-factorized matrix inverses in this context may be less visible, but it starts to shape the field of contemporary preconditioning techniques as well. One example may be the recent development of inverse-based dropping and related pivoting strategies [1].

Contemporary implementations of incomplete decompositions are typically improved by various enhancements like by permutations which enforce stronger diagonal dominance or by sophisticated modifications of matrix entries. Nevertheless, many of such approaches may improve the preconditioners just quantitatively and not qualitatively. The quantitative improvement may be very important from the practical point of view but it does not answer the important question whether we are able to improve incomplete decompositions adaptively by understanding the complex interplay between the direct decomposition and its inverse counterpart. Note that these two types of decomposition do not differ locally so much as we could suppose. For example, it is well-known that the corresponding direct and inverse elementary elimination matrices differ just by the sign of the off-diagonal entry and the results discussed below are heavily based on this fact. We believe that it is possible to understand the complementary role of both of these decompositions and use this knowledge to develop more robust algebraic preconditioners.

Our presentation will be mainly based on the two papers [2], [3] devoted to the orthogonalization techniques with a non-standard inner product. In our case we consider the inner product with the matrix \((D - A^{-1})^{-1}\), where \(A\) is the system matrix of the considered system of linear algebraic equations and \(D\) is a chosen auxiliary diagonal matrix. Such inner product puts together both direct and inverse elementary elimination matrices in each step of the decomposition. It was shown that this orthogonalization leads to the simultaneous computation of the direct and inverse decomposition. All the computed factors are approximate if the decomposition is incomplete. Practical computational procedure based on this technique connects the computation of the direct and inverse incomplete factors by dropping of some evaluated entries such that the actual dropping in the direct factor is directed by the growth of the inverse factor and vice versa. The paper [3] shows that it is possible to influence mutually the computation of both factors even without the inverse-based dropping. Nevertheless, purely theoretical considerations of this kind should be followed by practical preconditioning procedures. That is, they should take into account the possibility to

237
apply the theoretical rules inside practical implementations. The implementation should not have excessive memory demands and, at the same time, memory restrictions should not pose difficulties for fast access to the partially computed quantities which are used in the orthogonalization.

Our implementation clearly shows that it is possible to enhance the robustness of the incomplete decompositions in this way. The resulting procedure is cost-efficient and we believe that it can be used not only independently, but it can be a useful building block, for example, in more complex solvers. Nevertheless, a significant number of open questions remains and we would like to discuss at least some of them. First, there is a problem whether this approach can be efficiently combined with some kind of pivoting in order to provide even more powerful preconditioners. If this would be feasible then we could think, for example, about coupling of the new approach with the multilevel framework. Another problem connected to our search for efficient algebraic solvers can be posed as follows: we ask whether it is possible to compute a direct incomplete Cholesky decomposition similarly to the computation of the stabilized factorized inverse SAINV, namely via orthogonal projections. The question is motivated by the existence of the breakdown-free inverse incomplete decomposition based on the generalized modified Gram-Schmidt algorithm which performs reasonably well within the class of inverse decompositions. The answer is positive, but the algorithmic scheme is not straightforward since it includes some quantities which may be available only approximately. Our talk will include some of these developments including also applications of the discussed direct-inverse decomposition for condition estimation based on the joint work with Jurjen Duintjer Tebbens.


Iterative Solvers for Stochastic Galerkin Discretizations of the Lognormal Diffusion Problem

Elisabeth Ullmann

Abstract

Many physical processes occurring in different areas of science, engineering and industry are modelled by partial differential equations (PDEs). The numerical simulation of such processes requires input data which are, however, often subject to considerable uncertainty. Quantitative statements on the effect of these data uncertainties are therefore desirable for the evaluation of simulation results.

Stochastic Galerkin methods [1] are a well-established discretization tool for PDEs with uncertain data modelled in terms of random fields. The method couples physical degrees of freedom arising from standard finite element discretizations with stochastic degrees of freedom. For this reason the number of unknowns in the Galerkin equations grows rapidly: stochastic Galerkin equations can involve up to 10,000 times more unknowns than deterministic Galerkin equations. Consequently, the design of efficient and robust iterative solvers for these huge linear systems of equations is essential for the numerical simulation with uncertain data.

In recent years mostly two types of stochastic shape functions have been studied: multivariate orthogonal polynomials (spectral stochastic finite element method) and interpolation polynomials associated with certain quadrature nodes (stochastic collocation method). In this presentation we focus on the first approach, which precludes the decoupling of the Galerkin equations in general (see [4]).

Specifically, we are concerned with a model problem arising from steady-state groundwater flow calculations where the permeability of the rock is often modelled as a lognormal random field. Mathematically, this problem can be formulated as diffusion equation where the diffusion coefficient is a nonlinear function of a fixed number of statistically independent Gaussian random variables. We consider primal and mixed stochastic Galerkin finite element discretizations of our model problem and present an overview of iterative solution strategies. In particular, we discuss structural and spectral properties of the associated Galerkin matrices and outline the resulting computational challenges together with proposed preconditioners for the Galerkin system (see [2], [3]).

References

An Elegant IDR(s) Variant that Efficiently Exploits Bi-orthogonality Properties

Martin B. van Gijzen, Peter Sonneveld, and Tijmen P. Collignon

Abstract

The IDR(s) method [1] is a very efficient limited-memory method for solving large nonsymmetric systems of linear equations

\[ Ax = b, \]

with \( A \in \mathbb{C}^{N \times N} \), and \( x, b \in \mathbb{C}^N \). The method generates residuals that are forced to be in subspaces \( G_j \)'s related by \( G_j = (I - \omega_j A)(S \cap G_{j-1}) \), where \( S \) is a fixed proper subspace of \( \mathbb{C}^N \), and the \( \omega_j \)'s are non-zero scalars. According to the Induced Dimension Reduction theorem, on which IDR(s) is based, the subspaces \( G_j \)'s are nested and of strictly decreasing dimension. Without loss of generality, we can define \( S \) as the left null space of some \( N \times s \) matrix \( P \).

The prototype IDR(s) algorithm presented in [1] is a direct translation of the theorem into an algorithm and is only one of many possible variants. One of the possibilities for deriving a new IDR method is to choose a different way of computing the so-called intermediate residuals. In IDR(s), the residuals are uniquely defined at every \( s + 1 \)-st step. Such a step corresponds to the calculation of the first residual in a subspace \( G_j \). In order to advance to a new subspace \( G_{j+1} \), \( s \) additional (intermediate) residuals in the current subspace \( G_j \) need to be computed. The intermediate residuals are not uniquely defined. In exact arithmetic, the residuals at every \( s + 1 \)-st step are independent of the specific choice of the intermediate residuals. This property is lost in finite-precision computations. In this case, the numerical stability properties may severely influence the results, and therefore the behaviour in practice of the specific IDR method will depend on the intermediate residuals.

We will present an elegant, efficient, and, in our experience, numerically very stable IDR-based method that imposes and exploits as much as possible (one-sided) bi-orthogonality conditions between the intermediate residuals and the columns of \( P \). We denote this new IDR variant by IDR(s)-biortho to distinguish it from IDR(s)-proto, the prototype algorithm in [1]. IDR(s)-biortho uses fewer vector operations per iteration than IDR(s)-proto and has better stability properties, especially for large values of \( s \).

IDR(s)-biortho is also well suited for parallel computing. This is due to two facts. First, the algorithm can be formulated in such a way that there is only one global synchronization point per iteration. Second, it is possible to choose the IDR parameters \( s \) and \( P \) a priori to optimally tune the algorithm for any given parallel computer. The optimal settings depends only on some hardware parameters and on the problem size.

We will demonstrate the stability and efficiency of IDR(s)-biortho for a wide range of problems. Extensive comparisons of IDR(s)-biortho with state-of-the-art iterative and direct solution methods show that this new variant of IDR(s) is among the best solvers available.

We will also illustrate the parallel performance of the algorithm on a grid computer with geographically separated clusters.
A Matlab implementation of IDR(s)-biortho can be downloaded from the IDR web site: http://ta.twi.tudelft.nl/nw/users/gijzen/IDR.html. The code implements many advanced features and includes a test set and a manual that describes the functionality and illustrates the effects of different parameter settings. Also available from the web site are test problems and drivers for comparing the performance of IDR(s) with standard iterative solvers such as GMRES and Bi-CGSTAB.

The IDR(s)-biortho algorithm is described in detail in [2]. This report also describes extensive numerical experiments with the Matlab implementation, including comparisons with other methods. The parallelization and implementation on a grid computer are described in [3]. Both reports are available on the IDR web site.

References


Successful Tensor Decompositions in Clinical Practice

Sabine Van Huffel

Abstract

This contribution deals with tensor decompositions and their benefits in biomedical signal processing. In particular, we will focus on Parallel Factor Analysis (Parafac, also known as Candecomp or the CP model), the most popular tensor decomposition, and overview its mathematical properties and algorithms. The CP model decomposes in a unique way a higher-order tensor in a minimal sum of rank-1 atoms. Furthermore, we will give an overview of biomedical applications of these algorithms and their benefits will be illustrated in a variety of case studies in clinical practice. In particular, we will focus on artefact removal on Event-related potential (ERP) data from a visual detection task during simultaneous fMRI acquisition and on the localization of the onset of an epileptic seizure. This is joint work with Lieven De Lathauwer, Wouter Deburchgraeve and Maarten De Vos. More information is included in:

Block Tensor Computation

Charles Van Loan and Stefan Ragnarsson

Abstract

Will blocking become as important in tensor computations as it is in matrix computations? We will present results that address this question from three perspectives.

Blocking for Performance. A tensor contraction can be framed as a (typically large) collection of matrix multiplications. For example, if \( A \in \mathbb{R}^{N_1 \times N_1 \times N_2 \times N_2} \), \( B \in \mathbb{R}^{N_1 \times N_1 \times N_2 \times N_2} \), and

\[
C(i_1, j_2, i_3, i_4, j_3, j_4) = \sum_{k=1}^{N_1} A(i_1, k, i_3, i_4) B(k, j_2, j_3, j_4).
\]

then

\[
C(:, :, i_3, i_4, j_3, j_4) = A(:, :, i_3, i_4) B(:, :, j_3, j_4)
\]

This sample contraction involves \( N_2^4 \) matrix multiplications (MMs), each of which involves a pair of \( N_1 \)-by-\( N_1 \) matrices. A handy way to express the collection of the required MMs is via a block outer product. For example, if we set \( C_{ij\ell} = C(:, :, i, j, k, \ell) \), \( A_{i,j} = A(:, :, i, j) \), \( B_{ij} = B(:, :, i, j) \), then in the \( N_2 = 2 \) case we have

\[
\begin{bmatrix}
C_{1,1,1,1} & C_{1,1,1,2} & C_{1,1,2,1} & C_{1,1,2,2} \\
C_{2,1,1,1} & C_{2,1,1,2} & C_{2,1,2,1} & C_{2,1,2,2} \\
C_{1,2,1,1} & C_{1,2,1,2} & C_{1,2,2,1} & C_{1,2,2,2} \\
C_{2,2,1,1} & C_{2,2,1,2} & C_{2,2,2,1} & C_{2,2,2,2}
\end{bmatrix}
= \begin{bmatrix} A_{1,1} \\ A_{2,1} \\ A_{1,2} \\ A_{2,2} \end{bmatrix}
\begin{bmatrix} B_{1,1} & B_{2,1} & B_{1,2} & B_{2,2} \end{bmatrix}
\tag{1}
\]

Notice that each matrix slice from \( A \) and \( B \) is involved in \( N_2^4 \) MMs.

The role of the Strassen procedure in this venue will be described as will the tensor analog of GEMM-based BLAS, a handy way of handling symmetry. Block-tensor data structures are important.

Blocking for Generalization. A modal unfolding of a tensor is based on a fiber blocking of the tensor. For example, if \( A \) is 4-by-3-by-2, then the mode-2 unfolding is given by

\[
A_{(2)} = \begin{bmatrix}
a_{111} & a_{211} & a_{311} & a_{411} & a_{112} & a_{212} & a_{312} & a_{412} \\
a_{121} & a_{221} & a_{321} & a_{421} & a_{122} & a_{222} & a_{322} & a_{422} \\
a_{131} & a_{231} & a_{331} & a_{431} & a_{132} & a_{232} & a_{332} & a_{432}
\end{bmatrix}.
\]

The columns are mode-2 fibers:

\[
A_{(2)} = [ A_{(1,:,1)} \ A_{(2,:,1)} \ A_{(3,:,1)} \ A_{(4,:,1)} \ A_{(1,:,2)} \ A_{(2,:,2)} \ A_{(3,:,2)} \ A_{(4,:,2)} ].
\]

Various tensor decompositions and algorithms revolve around this particular way of turning a tensor computation into a matrix computation.

We will show that other unfoldings can provide a richer commentary on the source tensor when various types of structure are present. For example, if \( A \) is \( N \)-by-\( N \)-by-\( N \)-by-\( N \) with the property that

\[
A(i_1, i_2, i_3, i_4) = A(i_2, i_1, i_3, i_4) = A(i_1, i_2, i_4, i_3) = A(i_3, i_4, i_1, i_2)
\]

243
can be flattened into a symmetric block matrix $A$ whose blocks are symmetric and which satisfies $PAP^T = A$ where $P$ is the perfect shuffle. Such a matrix has a very special Kronecker product SVD which can be reshaped into an interesting tensor decomposition of $A$.

Blocking for Insight. The symmetric Schur decomposition of

$$\text{sym}(A) = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}$$

is highly related to the singular value decomposition of $A$. This connection has theoretical and computational ramifications. We have extended the $\text{sym}$ operator to tensors. The third-order case $C = \text{sym}(A)$ can be visualized as a 3-by-3-by-3 “Rubik cube” with six strategically positioned blocks that house the six transpositions of the source tensor:

We show that methods designed to compute $C$’s variationally-defined eigenvalues are equivalent to methods that can be applied to compute $A$’s variationally-defined singular values. Interesting open questions that surround the rank of $C$ as it relates to the rank of $A$ will also be presented.

Reference

S. Ragnarsson and C. Van Loan. “Block Tensors and Symmetric Embeddings,” submitted to the special tensor issue of *Linear Algebra and Its Applications.*
The Interplay of Givens Rotations and the $QR$-factorization in Inversion, Unitary Similarity Transforms and $QR$-algorithms

Raf Vandebril

Abstract

Givens rotations are already for decades a simple though valuable and reliable tool for annihilating specific elements in matrices. Their reliability and specific dedicated task makes them appear in a wide variety of numerical linear algebra algorithms. The easiness of understanding and implementing the transformations, the fact that they only alter specifically chosen rows and/or columns in a matrix makes them widespread and one of the basic building blocks in matrix computations.

In this overview we will study groups of Givens rotations and in particular their mutual interaction will be investigated. Since these transformation act only specified matrix parts, we need to take this into consideration. Graphical schemes will be presented containing all the vital information. These schemes will plainly visibilize the hopping around of the transformations, displaying as well the rows/columns affected by them.

One of the central ideas in this talk is the factorization of an arbitrary unitary matrix in specific patterns of rotations. This enables us to exploit all knowledge related to rotations for dealing with generic unitary matrices. For an arbitrary matrix $A$, the $QR$-factorization will be used. The $Q$-factor is decomposed in sequences of rotations and the interplay between rotations and the remaining upper triangular matrix $R$ is investigated. An almost trivial lemma enables us to transfer rotations through upper triangular matrices altering thereby the parameters determining the rotation itself, but not the other specifications such as the rows it acts on. These tools make it possible to reconsider well-known algorithms and to investigate them by focusing on the rotations involved, instead of on the original matrix.

Studying the structure of matrix inverses becomes an uncomplicated matter using the $QR$-factorization and the involved rotations. Reexamining the unitary similarity transformation to Hessenberg form reveals new insights and unveils the existence of related similarity transformations (to obtain e.g. a CMV-like decomposition of the unitary matrix). Finally the traditional $QR$-algorithms for Hessenberg and semiseparable matrices are subjected to a closer look from another viewpoint. It will be shown that both algorithms are instances of the same generic implicit rotation chasing idea.

If time permits attention will be paid to the polynomial root-finding problem. It will be shown how the associated companion and comrade matrices are examples of the more general unitary and Hermitian plus low rank eigenvalue problems. An implicit $QR$-approach for computing eigenvalues of both matrices will be presented based extensively on the tools presented before.
Riemannian and Multilevel Optimization for Rank-constrained Matrix Problems

Bart Vandereycken

Abstract

We present a Riemannian framework for the numerical solution of certain rank-constrained matrix problems. Let \( f : \mathbb{R}^{n \times n} \rightarrow \mathbb{R} \) be a smooth objective function defined on \( \mathbb{R}^{n \times n} \), the set of \( n \)-by-\( n \) real matrices. Then, these problems are of the form

\[
\min f(x) \quad \text{subject to} \quad x \in S^n_+,
\]

where \( S^n_+ \) denotes the \( n \)-by-\( n \) symmetric and positive semidefinite matrices of rank \( p \). Exploiting that the constrained set \( S^n_+ \) is a smooth, Riemannian manifold, enables us to solve (1) using techniques from smooth Riemannian optimization [3, 2].

The application of Riemannian optimization to \( S^n_+ \), however, requires some typical objects from differential geometry, like the tangent space and the Levi–Civita connection. As a first contribution, we therefore derive these objects for a submanifold geometry embedded in the Euclidean space. The geometry is chosen such that the geometrical objects can be implemented efficiently, that is, in \( O(np^2) \) for some small constant \( c \). In addition, we also derive and study the geodesics of this space.

The efficiency of the proposed framework is then assessed by solving large-scale matrix equations that originate from discretized partial differential equations (PDEs). Specifically, the goal is to compute a low-rank approximation of a matrix \( X \) that solves the Lyapunov equation,

\[
AXM^T + MXA^T = C,
\]

where \( A \) is the system matrix of the PDE, \( M \) the mass matrix, and \( C \) a known right-hand side matrix. In many applications \( C \) is such that \( X \) admits a good low-rank approximation; for example, \( C \) can be a low-rank matrix itself.

We propose two novel numerical methods. The first algorithm is the application of the Riemannian Trust-Region (RTR) method from [1] to a specifically chosen objective function in combination with the earlier derived submanifold geometry of \( S^n_+ \). In order to obtain an algorithm that is scalable for realistic PDEs, we derive a preconditioner that allows us to solve the Trust-Region subproblems of the RTR method much faster. Numerical experiments indicate that this algorithm is competitive compared to the state-of-the-art for low-rank Lyapunov solvers, while, at the same, it is more general and flexible.

The other algorithm is the generalization of multilevel optimization [4] to Riemannian manifolds. Compared to the previous solver, this multilevel Riemannian algorithm exploits the multilevel character of PDEs directly. This avoids the need for deriving a specific preconditioner, which makes this solver more flexible. However, in order to obtain an efficient multigrid solver, the typical multigrid components, like the smoother and the coarse grid operator, have to be chosen complementary. We show by a Local Fourier Analysis that the usual multigrid practice can be used for choosing these components. The numerical experiments illustrate that when the components are chosen in a correct way, one can indeed achieve textbook multigrid efficiency.
References


Finding Approximately Rank-one Submatrices with the Nuclear Norm and ℓ₁-norm

Xuan Vinh Doan and Stephen A. Vavasis

Abstract

Given a nonnegative matrix $A \in \mathbb{R}^{m \times n}$, we consider the problem of finding $I \subset \{1, \ldots, m\}$ and $J \subset \{1, \ldots, n\}$ such that $A(I, J)$ is close to a rank-one matrix and such that $\|A(I, J)\|$ is large. We shall call this problem the LAROS problem (for “large approximately rank-one submatrix”).

The main application of the LAROS problem is finding a feature in data. For example, suppose $A$ represents a corpus of documents in some language. Each column of $A$ is in correspondence with one document, and each row is in correspondence with a term used in the corpus. Here, “term” means a word in the language, excluding common words such as articles and prepositions. The $(i,j)$ entry of $A$ is the number of occurrences of term $i$ in document $j$, perhaps normalized. Such a matrix is called the term-document matrix of the underlying corpus. In this case, an approximately rank-one submatrix of $A$ corresponds to a subset of terms and a subset of documents in which the selected terms occur with proportional frequencies in the selected documents. Finding features using a LAROS formulation occurs as a subproblem in some algorithms for nonnegative matrix factorization, e.g., [1, 2, 4].

We propose the following formulation of LAROS: Find $X \in \mathbb{R}^{m \times n}$, $I \subset \{1, \ldots, m\}$, $J \subset \{1, \ldots, n\}$, solving

$$\min_{X} \|X\|_* + \theta \cdot |I| \cdot |J|$$

subject to $A \cdot X \geq 1$,

$$X(i,j) = 0 \text{ for } (i,j) \notin I \times J. \quad (1)$$

Here $\theta > 0$ is a penalty parameter to encourage a smaller matrix to be selected. The notation $A \cdot X$ denotes the inner product on $\mathbb{R}^{m \times n}$ and $\|\cdot\|_*$ is the so-called nuclear or trace norm and is defined as the sum of the singular values. The rationale of this formula is that, for a fixed choice of $I, J$, the optimal choice of $X$ according to duality theorems is $X(I, J) = U(:, 1)V(:, 1)^T / \Sigma(1, 1)$, where $U \Sigma V^T$ is the SVD of $A(I, J)$. Therefore, once $I, J$ are fixed, $X$ is selected to be the best rank-one approximation (up to a scale factor) to $A(I, J)$, and the submatrix with the largest first singular value has the optimal objective value. If $\theta = 0$, then the optimizer of (1) is the dominant singular-vector approximation to the entire matrix $A$. In some cases this answer may be undesirable as it can cause several distinct topics to be “averaged” together, as argued by [2]. For $\theta \rightarrow \infty$, the above formulation extracts a $1 \times 1$ submatrix corresponding to the largest absolute entry of $A$.

This formulation (1) is nonconvex and NP-hard in general, so we relax it to the following convex program:

$$\min_{X} \|X\|_* + \theta \|X\|_{\ell_1}$$

subject to $A \cdot X \geq 1. \quad (2)$

Here, $\|\cdot\|_{\ell_1}$ denotes the $\ell_1$-norm applied to vec$(X)$, or equivalently, the sum of the absolute values of the entries of $X$. The use of the $\ell_1$-norm as a convex relaxation of the number of nonzeros is a fundamental idea in the “compressed sensing” literature. We note that an objective function involving a sum of the nuclear and $\ell_1$-norms was used for a different purpose by [3]. Problem (2) is a convex programming problem and may be reformulated as semidefinite programming. We can prove the following theorem showing that it recovers approximate rank-one submatrices hidden by noise.
First, we recall the following definition: a random variable $x$ is $b$-subgaussian if its mean is zero, and if there exists a $b > 0$ such that for all $t \geq 0$, $\text{Prob}(|x| \geq t) \leq \exp(-t^2/(2b^2))$. For example, a normally distributed variable or any discretely distributed variable whose mean is zero is subgaussian.

Our main result is as follows.

**Theorem.** Let $A$ be an $m \times n$ matrix defined as follows.

$$A = \left( \begin{array}{cc} \sigma_0 u_0 v_0^T & 0 \\ 0 & 0 \end{array} \right) + \left( \begin{array}{cc} R_{11} & R_{12} \\ R_{21} & R_{22} \end{array} \right),$$

where $\sigma_0 > 0$, $u_0 \in \mathbb{R}^M$, $u_0 \geq 0$, $M < m$, and $v_0 \in \mathbb{R}^N$, $v_0 \geq 0$, $N < n$. Furthermore, assume that $u_0 = e_M + p$ with $\|p\| \leq c_1 \sqrt{M}$, and $v_0 = e_N + q$ with $\|q\| \leq c_2 \sqrt{N}$, where $e_M$, $e_N$ are all-1’s vectors of length $M$, $N$ respectively. The matrix $R$ is a random matrix with i.i.d. nonnegative elements $r_{ij}$ with mean $c_3 \sigma_0$, where $c_3 > 0$ is a constant, such that $r_{ij}/\sigma_0 - c_3$ is $b$-subgaussian. Here $c_1, c_2, c_3$ are positive constants. Assume $\theta$ satisfies $\sqrt{MN} \cdot \theta \in [\rho, \sigma]$, where $\rho, \sigma$ are scalars depending on $c_1, c_2, c_3$ (and $\rho < \sigma$ provided $c_1, c_2, c_3$ are sufficiently small). Then the solution $X$ to problem (2) is a rank-one matrix with positive entries in positions indexed by $\{1, \ldots, M\} \times \{1, \ldots, N\}$ and zeros elsewhere with probability exponentially close to 1 (i.e., of the form $1 - \exp(-(M + N)^\text{const})$) provided that $MN \geq \Omega((M + N)^{4/3})$ and $MN \geq \Omega(mn)$. Here, the constants implicit in the $\Omega(\cdot)$ notation depend on $b$ and $c_1, c_2$. See the full paper for a detailed presentation of these constants.

**Remarks.**

1. Naturally, the theorem also applies if the $MN$ distinguished entries occur as any $M \times N$ submatrix of $A$; we have numbered the distinguished submatrix first in order to simplify notation.

2. It is not enough to assume merely that $u_0 > 0$ and $v_0 > 0$ because if these vectors have very small entries, then they cannot be distinguished from the noise.

3. As a special case, observe that the result holds with probability exponentially close to 1 as long as $M \sim N$ and $M \geq \Omega(m^{1/2})$, $N \geq \Omega(n^{1/2})$. Thus, the relaxation can find approximately rank-one submatrices much smaller than the entire matrix $A$.

**References**


Preconditioned Locally Minimal Residual Method for Computing Interior Eigenpairs of Symmetric Operators

Eugene Vecharynski, Andrew Knyazev

Abstract

We consider the (generalized) symmetric eigenvalue problem

$$Av = \lambda Bv, \ A = A^*, \ B = B^* > 0,$$

where the targeted eigenpair corresponds to the smallest, in the absolute value, eigenvalue of the matrix pencil $A - \lambda B$. Large problems of this type frequently appear in applications, e.g., in the electronic structure calculations, where a number of eigenpairs of a Hamiltonian matrix around a given energy level need to be found.

We start by considering iterative methods for solving symmetric indefinite systems of linear equations with symmetric positive definite (SPD) preconditions. We briefly describe a hierarchy of such methods, from a stationary iteration to the optimal Krylov subspace preconditioned minimal residual method (PMINRES), and suggest a preconditioning strategy based on the approximation, in a certain sense, of the inverse of the absolute value of the coefficient matrix (absolute value preconditioners). We present an example of a simple (geometric) multigrid absolute value preconditioner for the symmetric model problem of the discretized real Helmholtz (shifted Laplacian) equation in two spatial dimensions with a relatively low wavenumber.

Next we extend the ideas underlying the SPD-preconditioned residual-minimizing methods for solving symmetric indefinite linear systems to the problem of computing an interior eigenpair of a (generalized) symmetric matrix pencil. We present a method that we call the preconditioned locally minimal residual method (PLMR), which represents a technique for finding an eigenpair corresponding to the smallest, in the absolute value, eigenvalue of a $A - \lambda B$. The method is based on the idea of the refined Rayleigh-Ritz procedure, in the preconditioner-based inner product, over four-dimensional trial subspaces, and relies on the choice of the (SPD) absolute value preconditioner. Importantly, the absolute value preconditioner for the eigenvalue problem can be the same as the one for the corresponding symmetric indefinite linear system with the coefficient matrix $A$. We illustrate the performance of the method on the example of the model problem of finding an interior eigenpair, which is the closest to a given shift value, of the two-dimensional Laplacian. We use the same multigrid preconditioner as for the Helmholtz equation mentioned above.
Spectral Intervals for Differential-algebraic Equations and Their Numerical Approximation by QR and SVD Algorithms

Vu Hoang Linh, Volker Mehrmann, and Erik Van Vleck

Abstract

In this talk, we discuss numerical methods for approximating spectral intervals for differential-algebraic equations (DAEs). Lyapunov and Sacker-Sell spectral intervals that are well-known for ordinary differential equations (ODEs) are extended to linear differential-algebraic equations with variable coefficients. For both the analytical and the numerical investigations, we use the strangeness index approach. Using any smooth basis of the solution subspace for a DAE and an appropriately chosen scaling factor, the so-called essentially underlying implicit ODE, which possesses the same spectral properties as the original DAE, can be constructed. For approximating the spectral intervals and their associated leading directions, numerical methods based on smooth matrix decompositions such as smooth QR and smooth SVD are proposed, which are applied directly to DAEs given in the strangeness-free form. We discuss different variants of these methods, their implementation, as well as the associated error analysis. Illustrative numerical experiments are given to confirm the efficiency of the methods.
Recursively Deflated PCG for Mechanical Problems

C. Vuik, T.B. Jönsthövel, and M.B. van Gijzen

Abstract

Introduction

We consider the simulation of displacements under loading for mechanical problems. Large discontinuities in material properties, such as encountered in composite materials, lead to ill-conditioned systems of linear equations. These discontinuities give rise to small eigenvalues that may negatively affect the convergence of the Preconditioned Conjugate Gradient (PCG) method.

This paper considers the Recursively Deflated Preconditioned Conjugate Gradient (RDPCG) method for solving such systems. Our deflation technique uses as deflation space the rigid body modes of sets of elements with homogeneous material properties. We show that in the deflated spectrum the small eigenvalues are mapped to zero and no longer negatively affect the convergence. We justify our approach through mathematical analysis and we show with numerical experiments on both academic and realistic test problems that the convergence of our RDPCG method is independent of discontinuities in the material properties. In many known examples there are only two materials involved: a weak and a stiff material. Our Recursively Deflated PCG method can be applied to a whole cascade of weaker and stiffer materials.

We consider asphalt concrete as an example of a composite material. It consists of a mixture of bitumen, aggregates and air voids. Obviously the difference between the stiffness of bitumen and the aggregates is significant, especially at high temperatures.

We simulate the response of a composite material that is subjected to external forces by means of small load steps. By using the FE method we obtain the corresponding stiffness matrix. Solving linear system $Ku = f$ is the most time consuming part of the FE simulation. The stiffness matrix $K$ is symmetric positive definite. We have shown in [1] that the number of iterations to convergence for PCG is highly dependent on the number of aggregates in a mixture as well as the ratio of the $E$ moduli. Increasing the number of aggregates introduces correspondingly more small eigenvalues in stiffness matrix $K$. The jumps in the $E$ moduli are related to the size of the small eigenvalues.

Recursively Deflated PCG

We define the projection $P$ by $P = I - KZ(Z^TKZ)^{-1}Z^T$ where $Z \in \mathbb{R}^{n \times m}$ is the deflation subspace. We assume that $m \ll n$ and $Z$ has rank $m$. Under this assumption $K_c \equiv Z^TKZ$ is symmetric positive definite and may be easily computed and factored. This definition does not provide insight in the effect of individual deflation vectors on the spectrum of $PK$. The next theorem defines a recursive deflation operator which can be used for more extensive eigenvalue analysis of $PK$.

**Definition 1.** \( P^{(k)} = I - KZ_k(Z^T_k KZ_k)^{-1}Z^T_k \) with $Z_k = [\tilde{Z}_1, \tilde{Z}_2, ..., \tilde{Z}_k]$, where $\tilde{Z}_j \in \mathbb{R}^{n \times l_j}$ and has rank $l_j$.

**Theorem 2.** Let $P^{(k)}$ and $Z_k$ as in Definition 1, then $P^{(k)}K = P_kP_{k-1} \cdots P_1K$ where $P_{i+1} = I - \tilde{K}_i\tilde{Z}_{i+1}(\tilde{Z}^T_{i+1}\tilde{K}_i\tilde{Z}_{i+1})^{-1}\tilde{Z}^T_{i+1}$, $\tilde{K}_i = P_i\tilde{K}_{i-1}$, $\tilde{K}_1 = P_1K$, $\tilde{K}_0 = K$, $\tilde{Z}^T_i\tilde{K}_{i-1}\tilde{Z}^T_i$ and $Z^T_iKZ_i$ are non-singular because $Z_i$ are of full rank and $K$ is a symmetric positive definite matrix.
Theorem 2 gives us a theoretical framework in which we construct the deflation vectors. We will see that by subsequently adding rigid body modes of particular sets of elements to the deflation space the number of small eigenvalues with deflation is smaller compared to without deflation.

**Rigid body modes as deflation vectors**

In this section we introduce a strategy to construct the deflation space $Z_j$ for $P^{(j)}K$ of Definition 1 to obtain decoupled problems. Our starting point is by observing that null spaces of sets of elements are represented by the rigid body modes of those sets of elements. There exists a splitting $K = C + R$ such that $C$ and $R$ are symmetric positive semi-definite with $\mathcal{N}(C) = \text{span}\{Z_k\}$ the null space of $C$. By choosing sets of elements we define $C$ and the nullspace of $C$ (our deflation space) is by definition spanned by the rigid body modes. We start with the elements which correspond to the most stiff material, followed by the lesser stiff materials, etc.

If some elements of a less stiff material are assigned to the element set of a stiffer material, the material stiffness matrices are not decoupled. We illustrate this with a simple example. When a node belongs to two elements and two different materials and is assigned to the wrong (least stiff) element with respect to the splitting of $K$, then the preconditioning step will reintroduce the coupling. For the 1D Poisson problem and preconditioning based on diagonal scaling, the entry on the main diagonal is $c_1 + c_2$, with $c_1 \ll c_2$. Clearly, when decoupled correctly, we have in splitting of $K$ only $c_2$ on the main diagonal of $C$, hence $L^{-1}CL^{-T}$ gives $\frac{c_2}{c_1 + c_2} \approx 1$. With a wrong choice of deflation vectors, we have $c_1$ on the main diagonal of $C$, hence $L^{-1}CL^{-T}$ gives $\frac{c_1}{c_1 + c_2} \approx \frac{1}{c_2} \ll 1$. However all other terms on the diagonal of $L^{-1}CL^{-T}$ will be approximately 1, introducing small eigenvalues for this material and unfavorable local spectrum of eigenvalues of $L^{-1}CL^{-T}$.

We consider a FE mesh of a real life sample of asphaltic material obtained from CT scan. We compare RDPCG and PCG in combination with incomplete Cholesky with drop tolerance $10^{-2}$. Table 1 shows that deflation leads to much less iterations and CPU time for various test problems.

<table>
<thead>
<tr>
<th></th>
<th>PCG iter</th>
<th>CPU (s)</th>
<th>RDPCG iter</th>
<th>CPU(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>i.</td>
<td>648</td>
<td>13.18</td>
<td>261</td>
<td>7.26</td>
</tr>
<tr>
<td>ii.</td>
<td>821</td>
<td>17.48</td>
<td>332</td>
<td>9.31</td>
</tr>
<tr>
<td>iii</td>
<td>756</td>
<td>15.21</td>
<td>331</td>
<td>8.89</td>
</tr>
</tbody>
</table>

**References**

Anderson Acceleration for Fixed-Point Iterations

Homer F. Walker and Peng Ni

Abstract

Fixed-point iterations occur naturally and are commonly used in a broad variety of computational science and engineering applications. In practice, fixed-point iterates often converge undesirably slowly, if at all, and procedures for accelerating the convergence would be useful. This talk will focus on a particular acceleration method that was originally proposed by D. G. Anderson in 1965 and will be referred to as Anderson acceleration here. This method has enjoyed considerable success and wide usage in electronic structure computations, where it is known as Anderson mixing; however, it seems to have been untried or underexploited in many other important applications. Moreover, while other acceleration methods have been extensively studied by the mathematics and numerical analysis communities, this method has received almost no attention from these communities over the years, with two notable exceptions. First, the method was independently rediscovered by K. Miller as early as 1990 and since applied by him and his collaborators to the numerical solution of PDEs in several contexts. Second, in a 2009 paper, H. Fang and Y. Saad have clarified a remarkable relationship of Anderson acceleration to quasi-Newton (secant updating) methods and extended it to define a broader Anderson family of acceleration methods.

The goals of this talk are to shed additional light on Anderson acceleration and to draw further attention to its usefulness as a general tool. The first part of the talk will cover theoretical results. The main result is that, on linear problems, Anderson acceleration without truncation is “essentially equivalent” in a certain sense to the generalized minimal residual (GMRES) method. A corollary is that applying Anderson acceleration to a classical stationary iteration for solving $Ax = b$ based on a splitting $A = M - N$ is “essentially equivalent” in the same sense to applying GMRES to the left-preconditioned system $M^{-1}Ax = M^{-1}b$. (However, Anderson acceleration may break down or suffer ill-conditioning before the solution is reached, and so applying it in this way is not recommended as a general alternative to preconditioned GMRES.) Counterparts of these results will be outlined showing analogous relationships between the Type 1 variant in the Fang–Saad Anderson family and the Arnoldi (FOM) method. The second part of the talk will address practical considerations for implementing Anderson acceleration and illustrate its performance through numerical experiments involving a variety of applications, including the EM algorithm for estimating statistical mixture parameters, alternating non-negative least-squares for non-negative matrix factorization, and domain-decomposition iterations for linear and nonlinear PDE problems.
The Next Step in the Never-ending Process of Generalizing Francis’s Implicitly-shifted $QR$ Algorithm

Raf Vandebril and David S. Watkins

Abstract

The first author recently proposed a large family of condensed forms that includes the Hessenberg, inverse-Hessenberg, and CMV forms as special cases. He showed how to obtain these forms and how to implement explicit and implicit $QR$-like algorithms on them. More recently we have shown how to implement implicit $QR$-like iterations of arbitrary degree, and we have developed the convergence theory of these algorithms. The condensed form evolves over the course of the iterations in a way that can be controlled. For example, Hessenberg form can be transformed to CMV or inverse-Hessenberg form over the course of many iterations. This capability of controlling the nature of the condensed form also allows some scope for manipulating the convergence rate, as we shall show.
Efficient Structured Solvers and Preconditioners for Large Sparse Linear Systems

Jianlin Xia, Ming Gu, and Xiaoye S. Li

Abstract

In recent years, it has been discovered that certain PDEs can be solved by direct factorization in nearly linear time. These techniques are closely related to domain decomposition, the fast multiple methods, multigrid, data compression, etc. A fundamental idea is a low-rank property. During the direct factorization of the discretized sparse matrix, the off-diagonal blocks of the dense Schur complements have small numerical ranks with a given tolerance. The Schur complements can then be approximated by rank structured matrices such as $\mathcal{H}$-matrices, hierarchically semiseparable (HSS) matrices, and quasiseparable matrices. This leads to the development of structured sparse direct solvers such as $\mathcal{H}$-LU methods [3, 4] and structured multifrontal methods [5].

These structured sparse methods generally assume that the off-diagonal numerical ranks of the intermediate dense matrices are reasonably bounded. Such a condition holds, say, for 2D elliptic equations, where the bound is $O(1)$ [1]. However, for other PDEs, especially 3D problems, these bounds maybe $O(N)$ or larger [1], where the mesh dimension is $N \times N \times N$.

We propose new structured methods with relaxed rank requirements based on new HSS algorithms and sparse matrix techniques such as nested dissection and the multifrontal method. The new HSS factorizations do not impose the classical strict rank assumption. Instead, we allow the off-diagonal ranks to increase as certain functions of the off-diagonal block (row) sizes. Such HSS algorithms are then used to partially factorize the frontal matrices in the multifrontal method for sparse matrices. Unlike the method in [5], we use a new data-sparse extend-add operation without complicated HSS permutations and splitting. This is dramatically more convenient to implement. The complexity of such a structured sparse factorization is listed in Table 1 together with the corresponding off-diagonal rank bounds of the intermediate Schur complements.

<table>
<thead>
<tr>
<th>Rank (at level $l$)</th>
<th>Rand bound</th>
<th>Factorization</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$O((\log N_i)^p)$ $(p \in \mathbf{N})$</td>
<td>$O((\log N)^p)$</td>
<td>$O(n) \sim O(n \log n)$</td>
<td>$O(n) \sim O(n \log \log n)$</td>
</tr>
<tr>
<td>$O(N_i^{1/p})$ $(p \in \mathbf{N}, p \geq 3)$</td>
<td>$O(N_i^{1/p})$</td>
<td>$O(n) \sim O(n \log n)$</td>
<td>$O(n) \sim O(n \log \log n)$</td>
</tr>
<tr>
<td>$O(N_i^{1/2})$</td>
<td>$O(N_i^{1/2})$</td>
<td>$O(n) \sim O(n \log n)$</td>
<td>$O(n) \sim O(n \log \log n)$</td>
</tr>
</tbody>
</table>

Table 1: Complexity of a structured multifrontal method with relaxed off-diagonal rank bounds for the intermediate dense Schur complements, where $n$ is the size of the matrix, and $N_i$ denotes the block row size of the level-$l$ HSS blocks of a Schur complement.
Our new method is significantly simpler and more flexible than the structured factorization method in [5]. The method does not require regular meshes. The efficiency of is illustrated with the aid of numerical experiments from various practical applications in areas such as seismic imaging.

The new structured factorization can also include robustness techniques and is suitable for preconditioning. One way is to use a fast factorization method to generate triangular HSS factors. The fundamental idea can be illustrated in terms of the factorization of a dense block 2 × 2 matrix

\[
A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} L_{11} \\ L_{22} \end{pmatrix} \begin{pmatrix} I & L_{11}^{-1} A_{12} L_{22}^{-T} \\ L_{22}^{-1} A_{21} L_{11}^{-T} & I - L_{11}^{-1} A_{12} L_{22}^{-T} \end{pmatrix} \begin{pmatrix} L_{11}^T \\ L_{22}^T \end{pmatrix}.
\] (1)

The off-diagonal block \( L_{22}^{-1} A_{21} L_{11}^{-1} \) is compressed with an approximate QR factorization

\[
L_{22}^{-1} A_{21} L_{11}^{-1} = \begin{pmatrix} U & \hat{U} \\ \Sigma & \hat{\Sigma} \end{pmatrix} \begin{pmatrix} V^T \\ \hat{V}^T \end{pmatrix} \approx U \Sigma V^T.
\]

Then continue to factorize the middle matrix on the right-hand side of (1). At this point, the Schur complement can be shown to be always positive definite, regardless of the accuracy in the compression of \( L_{22}^{-1} A_{21} L_{11}^{-1} \). This gives an approximate factorization \( \hat{L} \hat{L}^T \) which approximates \( A \) with an accuracy \( \sigma_{r+1} \), the largest singular value of \( \Sigma \). Furthermore, if \( \hat{L} \hat{L}^T \) is used to precondition \( A \), the preconditioned matrix has condition number \( \kappa_2(\hat{L}^{-1} A \hat{L}^{-T}) = 1 + \frac{2\sigma_{r+1}}{1-\sigma_{r+1}} \). Hence, \( \kappa_2(\hat{L}^{-1} A \hat{L}^{-T}) \) decays much faster than \( \sigma_{r+1} \), which indicates that even if \( A \) does not have a small off-diagonal numerical rank (weak low-rank property), we can still use a compact preconditioner \( \hat{L} \hat{L}^T \) to significantly improve the condition of \( A \).

Such an idea can be recursively generalized to multiple blocks with the aid of HSS matrices, using fast HSS solution methods. This is more general than the structured method in [6]. Such a dense factorization method can then be incorporated into the multifrontal method so as to precondition sparse problems. Numerical experiments indicate that, when preconditioning some PDEs, this method is significantly more effective than both some classical preconditioners and some other structured preconditioners.

References


Efficient Preconditioned Inner Solves for Inexact Rayleigh Quotient Iteration and Their Connections to the Simplified Jacobi-Davidson Method

Fei Xue, and Daniel B. Szyld

Abstract

We study inexact Rayleigh quotient iteration (RQI) for solving generalized eigenvalue problems, with focus on efficient preconditioned iterative solution of the shifted linear systems that arise in this algorithm. We give a new convergence analysis of inexact RQI, showing that locally cubic and quadratic convergence can be achieved for Hermitian and non-Hermitian problems, respectively, if the shifted linear systems are solved by a generic Krylov subspace method with a special "tuned" preconditioner to a reasonably small fixed tolerance. We then refine the study by Freitag and Spence [Linear Algebra Appl., Vol. 428 (2008), pp. 2049–2060] on the equivalence of the inner solves of inexact RQI and the simplified Jacobi-Davidson method where a preconditioned full orthogonalization method is used as the inner solver. We also provide some new perspectives on the tuning strategy, showing that tuning is essentially needed only in the first inner iteration in the non-Hermitian case. Two alternative inner solvers based on this observation are studied and compared. One of these alternatives, a flexible GMRES algorithm with a special configuration in the first inner step, can be comparable in efficiency to GMRES with the tuned preconditioner.
A Parallel Hybrid Linear Solver for Large-scale Highly-indefinite Linear Systems of Equations

Xiaoye Li, Esmond Ng, and Ichitaro Yamazaki

Abstract

Many forefront numerical simulations require solution of large-scale highly-indefinite linear systems of equations. The success of such simulations depends on the development of a robust and efficient linear solver which can effectively utilize thousands of processors and efficiently solve these linear systems. Great progress has been made in the development of high-performance parallel direct solvers which can robustly compute solutions of high accuracy. However, their performance typically stagnates on tens or hundreds of processors, and the matrices that can be directly factorized are limited in size due to large memory and communication requirements. Preconditioned iterative solvers require less memory and communication, but for the ill-conditioned highly-indefinite problems of our interest, they often require a high quality preconditioner which is not easily available.

To overcome these challenges, we have been developing a parallel hybrid linear solver which is based on a non-overlapping domain decomposition technique called the Schur complement method. In this method, the unknowns associated with each interior subdomains are first eliminated to form the Schur complement system, which is then solved using a preconditioned iterative solver. This method provides a framework for developing an effective parallel hybrid solver, where subdomain problems can be solved using a various approach including a parallel direct solver, a parallel preconditioned iterative solver, or even a hybrid solver. Furthermore, this framework provides two levels of parallelism: solving independent subdomain problems in parallel and using multiple processors per subdomain. This allows us to increase the processor count without increasing the number of subdomains or the size of the Schur complement, while requiring the solver on each subdomain to scale only up to tens or hundreds of processors. Hence, this hierarchical parallelism can lead to a scalable implementation which maintains numerical stability at the same time.

Even though this hybrid framework allows us to bring together much of the progress made in the development of linear solvers, their integration needs to be carefully designed. For instance, a parallel direct solver may be effective to factorize the interior subdomain using tens or hundreds of processors. However, in order to achieve high performance, it is imperative that the direct solver is extended to exploit the special structures of the subdomain problems arising from our software framework. For this, we developed several techniques such as taking advantage of the sparsity of right-hand-sides during the triangular solutions with interfaces, reordering the sparse right-hand-sides to improve the data locality, load balancing sparse matrix-matrix multiplication to form update matrices, and designing an effective asynchronous point-to-point communication of the update matrices. Furthermore, there are a number of matrix preprocessing techniques developed to enhance the performance of the iterative solver to solve highly-indefinite systems, which need to be carefully integrated to preprocess the Schur complement in parallel [4].

In this talk, we describe two applications which require the solution of large-scale highly-indefinite linear systems of equations; specifically, numerical simulations to model fusion devices [1] and those to design accelerator cavities [2]. Then, we discuss how we have combined and specialized the existing state-of-the-art techniques in order to efficiently solve these linear systems on thousands of processors. We present numerical results to demonstrate that our hybrid solver can efficiently solve a problem with 18 million unknowns from one of these simulations using more than 4,000 processors, whereas a parallel direct solver and many other iterative solvers fail.
References


Solving Nonlinear Eigenvalue Problems in Electronic Structure Calculation

Chao Yang, Juan C. Meza

Abstract

I will discuss numerical algorithms for solving discretized Kohn-Sham equations. This type of problems arises from electronic structure calculation which is nowadays an essential tool for studying the microscopic quantum mechanical properties of molecules, solids and other nanoscale materials. Through the density functional theory (DFT) formalism, one can reduce the many-body Schrödinger equation, which is a linear eigenvalue problem, used to describe the electron-electron and electron-nucleus interactions to a set of single-electron equations that have far fewer degrees of freedom. However, the price one has to pay is to solve a nonlinear eigenvalue problem in which the matrix Hamiltonian is a function of the desired but unknown eigenvectors to be computed. Currently, the most widely used numerical algorithm for solving this type of problem is the Self-consistent field (SCF) iteration. The algorithm can be viewed as a fixed point iteration in which approximations to the eigenvectors of a fixed Hamiltonian are computed at each iteration, and these eigenvectors are used to update the Hamiltonian for the next iteration. However, it is well known that the simplest form the SCF iteration often fails to converge to the correct solution. I will present an analysis of the SCF iteration that shows the conditions under which the SCF iteration forms a contraction mapping on the space of electron charge densities. If these conditions are satisfied, the SCF iteration is guaranteed to converge to the correct solution. I will discuss various strategies for improving the convergence of SCF when such conditions are violated. An alternative algorithm that is designed to minimize the total energy of the atomistic system of interest directly will also be presented. In this algorithm, the total energy of the system is minimized in a sequence of overlapping subspaces. The minimizer within each subspace is obtained by applying an enhanced SCF iteration to a projected nonlinear eigenvalue problem. I will demonstrate how this algorithm can be used to compute the electronic structures of both molecules and solids. Techniques for overcoming convergence difficulties in metallic systems and achieving linear scaling complexity will also be discussed.
A Fourier-Series-Based Kernel-Independent Fast Multipole Method

Bo Zhang

Abstract

We present a new kernel-independent fast multipole method, to which we refer as FKI-FMM, for efficient evaluation of the pairwise interactions governed by

$$u(x_i) = \sum_{j=1, j \neq i}^{N} q_j K(x_i - x_j), \quad i = 1, \ldots, N,$$

where the kernel function $K$ is translation invariant. This matrix-vector multiplication is used at every time step in large-scale simulation of molecular dynamics [1]. It is also a basic step in iterative solvers for certain discretized integral equations [2]. In these applications, the particles are not necessarily located on the points of an equally spaced grid. The interaction matrix $K(i, j) = K(x_i - x_j)$ consequently deviates substantially from the nested Toeplitz structure displayed by a translation-invariant kernel on equally-spaced grid points. This implies that the use of the fast Fourier transform (FFT) by the discrete convolution theorem does not apply directly. Fortunately, the fast multipole method (FMM) originally by Greengard and Rokhlin [3, 4] enabled the efficient evaluation in an alternative approach. It takes only $O(N)$ arithmetic operations instead of $O(N^2)$ as with the direct evaluation.

It has since been interesting and intimidating to study the relationship between the FMM and the discrete Fourier transform [5, 6, 7]. FKI-FMM has successfully exposed and exploited the relationship. It creates, using numerical techniques, a compressive representation of the interaction matrix, in sufficient accuracy, over multi-scale interaction regions in the form of a truncated Fourier series. It also provides, using analytical and algebraic approaches, the operators for the multipole-to-multipole, multipole-to-local, and local-to-local translations in the FMM framework. Among other advantages, the technique for compressive matrix representation is kernel independent and the matrix blocks involved in cross interactions at each and every spatial scale are readily diagonalized in the Fourier series modes. This leads to a greater scope of FMM-enabled applications, to further reduction in arithmetic operations for function evaluation as well as for matrix-vector product, and to substantial and important simplification in data, algorithm and programming structures.

FKI-FMM draws upon three major advances, since 1987, in the design, development, and applications of the FMM. These advances are on adaption to the particle distribution, diagonalization of multipole-to-local translation operators, and kernel-independent expansion mechanism. We demonstrate the differences FKI-FMM makes in kernel-independent matrix compression, in computation efficiency, as well as in construction of sequential and parallel algorithms. This is a joint work with J. Huang, N. P. Pitsianis and X. Sun.

References


List of Participants
<table>
<thead>
<tr>
<th>Name</th>
<th>Email</th>
<th>Institution and Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Al-Mohy, Awad H.</td>
<td><a href="mailto:aalmohy@hotmail.com">aalmohy@hotmail.com</a></td>
<td>King Khalid University, Saudi Arabia</td>
</tr>
<tr>
<td>Amsallem, David</td>
<td><a href="mailto:amsallem@stanford.edu">amsallem@stanford.edu</a></td>
<td>Stanford University, US</td>
</tr>
<tr>
<td>Antoulas, Athanasios C.</td>
<td><a href="mailto:aca@rice.edu">aca@rice.edu</a></td>
<td>Rice University, US</td>
</tr>
<tr>
<td>Arioli, Mario</td>
<td><a href="mailto:mario.arioli@stfc.ac.uk">mario.arioli@stfc.ac.uk</a></td>
<td>STFC Rutherford Appleton Laboratory, UK</td>
</tr>
<tr>
<td>Avron, Haim</td>
<td><a href="mailto:haimav@us.ibm.com">haimav@us.ibm.com</a></td>
<td>IBM T.J. Watson Research Center, US</td>
</tr>
<tr>
<td>Bai, Zhaojun</td>
<td><a href="mailto:bai@cs.ucdavis.edu">bai@cs.ucdavis.edu</a></td>
<td>University of California, Davis, US</td>
</tr>
<tr>
<td>Ballard, Grey</td>
<td><a href="mailto:ballard@cs.berkeley.edu">ballard@cs.berkeley.edu</a></td>
<td>University of California, Berkeley, US</td>
</tr>
<tr>
<td>Baum, Ann-Kristin</td>
<td><a href="mailto:baum@math.tu-berlin.de">baum@math.tu-berlin.de</a></td>
<td>TU Berlin, Germany</td>
</tr>
<tr>
<td>Beattie, Christopher</td>
<td><a href="mailto:beattie@vt.edu">beattie@vt.edu</a></td>
<td>Virginia Polytechnic Institute and State Uni-</td>
</tr>
<tr>
<td>Benner, Peter</td>
<td><a href="mailto:benner@mpi-magdeburg.mpg.de">benner@mpi-magdeburg.mpg.de</a></td>
<td>Max Planck Institute for Dynamics of Complex</td>
</tr>
<tr>
<td>Benzi, Michele</td>
<td><a href="mailto:benzi@mathcs.emory.edu">benzi@mathcs.emory.edu</a></td>
<td>Emory University, US</td>
</tr>
<tr>
<td>Bientinesi, Paolo</td>
<td><a href="mailto:pauldj@aices.rwth-aachen.de">pauldj@aices.rwth-aachen.de</a></td>
<td>RWTH Aachen University, Germany</td>
</tr>
<tr>
<td>Bindel, David</td>
<td><a href="mailto:bindel@cs.cornell.edu">bindel@cs.cornell.edu</a></td>
<td>Cornell University, US</td>
</tr>
<tr>
<td>Boley, Daniel</td>
<td><a href="mailto:boley@cs.umn.edu">boley@cs.umn.edu</a></td>
<td>University of Minnesota, US</td>
</tr>
<tr>
<td>Bora, Shreemayee</td>
<td><a href="mailto:shbora@iitg.ac.in">shbora@iitg.ac.in</a></td>
<td>Indian Institute of Technology Guwahati, In-</td>
</tr>
<tr>
<td>Borsdorf, Rüediger</td>
<td><a href="mailto:borsdorf@maths.man.ac.uk">borsdorf@maths.man.ac.uk</a></td>
<td>The University of Manchester, UK</td>
</tr>
<tr>
<td>Botchev, Mike A.</td>
<td><a href="mailto:m.a.botchev@math.utwente.nl">m.a.botchev@math.utwente.nl</a></td>
<td>University of Twente, The Netherlands</td>
</tr>
<tr>
<td>Braman, Karen</td>
<td><a href="mailto:Karen.Braman@sdsmt.edu">Karen.Braman@sdsmt.edu</a></td>
<td>South Dakota School of Mines and Technology,</td>
</tr>
<tr>
<td>Buttari, Alfredo</td>
<td><a href="mailto:alfredo.buttari@enseeiht.fr">alfredo.buttari@enseeiht.fr</a></td>
<td>CNRS-IRIT, France</td>
</tr>
<tr>
<td>Carden, Russell</td>
<td><a href="mailto:Russell.L.Carden@rice.edu">Russell.L.Carden@rice.edu</a></td>
<td>Rice University, US</td>
</tr>
</tbody>
</table>
Ernst, Oliver G.
ernst@math.tu-freiberg.de
TU Bergakademie Freiberg, Germany

Fassbender, Heike
h.fassbender@tu-braunschweig.de
TU Braunschweig, Germany

Fong, David Chin-lung
clfong@stanford.edu
Stanford University, US

Foster, Leslie V.
foster@math.sjsu.edu
San José State University, US

Freitag, Melina A.
m.freitag@maths.bath.ac.uk
University of Bath, UK

Friedland, Shmuel
friedlan@uic.edu
University of Illinois at Chicago, US

Frommer, Andreas
frommer@math.uni-wuppertal.de
Bergische Universität Wuppertal, Germany

Güttel, Stefan
guettel@gmail.com
University of Geneva, Switzerland

Gleich, David F.
dgleich@stanford.edu
Sandia National Laboratories, US

Gragg, Bill
bill_gragg@yahoo.com
Naval Postgraduate School, US

Greenbaum, Anne
greenbau@math.washington.edu
University of Washington, US

Greif, Chen
greif@cs.ubc.ca
The University of British Columbia, Canada

Grigori, Laura
Laura.Grigori@inria.fr
INRIA Saclay, Université Paris-Sud 11, France

Grubišić, Luka
luka@math.hr
University of Zagreb, Croatia

Gu, Ming
mgu@math.berkeley.edu
University of California, Berkeley, US

Gugercin, Serkan
gugercin@math.vt.edu
Virginia Polytechnic Institute and State University, US

Guo, Chun-Hua
chguo@math.uregina.ca
University of Regina, Canada

Gurbuzbalaban, Mert
mert@cims.nyu.edu
New York University, US

Gutknecht, Martin H.
mhg@math.ethz.ch
ETH Zürich, Switzerland

Higham, Nick
higham@maths.manchester.ac.uk
University of Manchester, UK

Hillar, Christopher J.
chillar@msri.org
MSRI: Mathematical Sciences Research Institute, US
Hnětynková, Iveta
hnetynkova@cs.cas.cz
Charles University in Prague, Czech Republic

Hoemmen, Mark
mhoemme@sandia.gov
Sandia National Laboratories, US

Huckle, Thomas
huckle@in.tum.de
TU Munich, Germany

Iannazzo, Bruno
bruno.iannazzo@dmi.unipg.it
Università di Perugia, Italy

Ipsen, Ilse
ipsen@ncsu.edu
North Carolina State University, US

Jarlebring, Elias
elias.jarlebring@cs.kuleuven.be
Katholieke Universiteit Leuven, Belgium

Jiránek, Pavel
pavel.jiranek@gmail.com
CERFACS, France

Johansson, Stefan
stefanj@cs.umu.se
Umeå University, Sweden

Kågström, Bo
bokg@cs.umu.se
Umeå University, Sweden

Kahan, William
wkahan@eecs.berkeley.edu
University of California, US

Karlsson, Lars
larvk@cs.umu.se
Umeå University and HPC2N, Sweden

Kaufman, Linda
kaufmanl@wpunj.edu
William Paterson University, US

Kilmer, Misha E.
misha.kilmer@tufts.edu
Tufts University, US

Kjelgaard Mikkelsen, Carl Christian
spock@cs.umu.se
Umeå University, Sweden

Kolda, Tamara G.
tgkolda@sandia.gov
Sandia National Laboratories, US

Kressner, Daniel
daniel.kressner@epfl.ch
EPF Lausanne, Switzerland

Langou, Julien
julien.langou@ucdenver.edu
University of Colorado Denver, US

Li, Ren-Cang
rcrl@uta.edu
University of Texas at Arlington, US

Li, Xiaoye S.
xсли@lbl.gov
Lawrence Berkeley National Laboratory, US

Liesen, Jörg
liesen@math.tu-berlin.de
TU Berlin, Germany

Lim, Lek-Heng
lekheng@galton.uchicago.edu
University of Chicago, US

Lin, Lijing
Lijing.Lin@postgrad.manchester.ac.uk
University of Manchester, UK
Loh, Po-Ru
ploh@mit.edu
MIT, US

Mackey, D. Steven
steve.mackey@wmich.edu
Western Michigan University, US

Mackey, Niloufer
nil.mackey@wmich.edu
Western Michigan University, US

Maes, Christopher
cmaes@mit.edu
MIT, US

Mahoney, Michael W.
mmahoney@theory.stanford.edu
Stanford University, US

Malyshev, Alexander
alexander.malyshev@math.uib.no
University of Bergen, Norway

Markovsky, Ivan
im@ecs.soton.ac.uk
University of Southampton, UK

Mastronardi, Nicola
n.mastronardi@ba.iac.cnr.it
National Research Council of Italy, Italy

Meerbergen, Karl
Karl.Meerbergen@cs.kuleuven.be
Katholieke Universiteit Leuven, Belgium

Mehl, Christian
mehl@math.tu-berlin.de
TU Berlin, Germany

Mehrmann, Volker
mehrmann@math.TU-Berlin.DE
TU Berlin, Germany

Mengi, Emre
emengi@ku.edu.tr
Koc University, Turkey

Miedlar, Agnieszka
miedlar@math.tu-berlin.de
TU Berlin, Germany

Molera, Juan M.
molera@math.uc3m.es
Universidad Carlos III de Madrid, Spain

Morgan, Ron
Ronald_Morgan@baylor.edu
Baylor University, US

Moro, Julio
jmoro@math.uc3m.es
Universidad Carlos III de Madrid, Spain

Nagy, James
nagy@mathcs.emory.edu
Emory University, US

Nakatsukasa, Yuji
ynakatsukasa@ucdavis.edu
University of California, Davis, US

Ng, Esmond G.
EGNg@lbl.gov
Lawrence Berkeley National Laboratory, US

O’Leary, Dianne P.
oleary@cs.umd.edu
University of Maryland, US
Oseledets, Ivan
ivan.oseledets@gmail.com
Institute of Numerical Mathematics, Russia

Overton, Michael L.
overton@cs.nyu.edu
New York University, US

Paige, Christopher
paige@cs.mcgill.ca
McGill University, Canada

Panayotov, Ivo
Ivo.Panayotov@maths.ox.ac.uk
University of Oxford, UK

Parlett, Beresford
parlett@math.berkeley.edu
University of California, Berkeley, US

Pestana, Jennifer
pestana@maths.ox.ac.uk
University of Oxford, UK

Plešinger, Martin
martin.plesinger@sam.math.ethz.ch
ETH Zürich, Switzerland

Poloni, Federico
f.poloni@sns.it
Scuola Normale Superiore, Italy

Ragnarsson, Stefan
str23@cornell.edu
Cornell University, US

Ramage, Alison
A.Ramage@strath.ac.uk
University of Strathclyde, UK

Rees, Tyrone
rees@math.ubc.ca
University of British Columbia, Canada

Renaut, Rosemary
rrenaut@nsf.gov
National Science Foundation/Arizona State University, US

Rojas, Marielba
marielba.rojas@tudelft.nl
Delft University of Technology, The Netherlands

Rozložník, Miroslav
miro@cs.cas.cz
Academy of Sciences of the Czech Republic, Czech Republic

Ruhe, Axel
ruhe@kth.se
Royal Institute of Technology, Sweden

Sadok, Hassane
sadok@lmpa.univ-littoral.fr
University of Littoral, France

Saunders, Michael
saunders@stanford.edu
Stanford University, US

Savas, Berkant
berkant@cs.utexas.edu
Linköping University, Sweden

Schröder, Christian
schroed@math.tu-berlin.de
TU Berlin, Germany

Schwartz, Oded
oded.schwartz@gmail.com
University of California, Berkeley, US

Scott, Jennifer
jennifer.scott@stfc.ac.uk
STFC Rutherford Appleton Laboratory, UK
Shah, Mili
mishah@loyola.edu
Loyola University Maryland, US

Sifuentes, Josef
sifuentes@courant.nyu.edu
New York University, US

Simoncini, Valeria
valeria@dm.unibo.it
Università di Bologna, Italy

G.L.G.Sleijpen@uu.nl

Soodhalter, Kirk M.
ksoodha@temple.edu
Temple University, US

Sorensen, Danny C.
sorensen@rice.edu
Rice University, US

Spence, Alastair
a.spence@bath.ac.uk
University of Bath, UK

Sra, Suvrit
suvrit.sra@tuebingen.mpg.de
Max Planck Institut für Biological Cybernetics, Germany

Stewart, G.W.
stewart@cs.umd.edu
University of Maryland, US

Stoll, Martin
martin.stoll80@gmail.com
Max Planck Institute for Dynamics of Complex Technical Systems, Germany

Strakoš, Zdeněk
strakos@cs.cas.cz
Charles University in Prague, Czech Republic

Sun, Xiaobai Sun
xiaobai@cs.duke.edu
Duke University, US

Sutton, Brian D.
briandavidsutton@gmail.com
Randolph-Macon College, US

Szyld, Daniel B.
szyld@temple.edu
Temple University, US

Tuňma, Miroslav
tuma@cs.cas.cz
Academy of Sciences of the Czech Republic, Czech Republic

Thorne, Sue
sue.thorne@stfc.ac.uk
STFC Rutherford Appleton Laboratory, UK

Tichý, Petr
tichy@cs.cas.cz
Academy of Sciences of the Czech Republic, Czech Republic

Tisseur, Françoise
ftisseur@ma.man.ac.uk
The University of Manchester, UK

Titeley-Peloquin, David
David.Titeley-Peloquin@maths.ox.ac.uk
University of Oxford, UK

Ullmann, Elisabeth
ullmann@math.tu-freiberg.de
TU Bergakademie Freiberg, Germany

Van Dooren, Paul
paul.vandooren@uclouvain.be
Université catholique de Louvain, Belgium
van Gijzen, Martin B. 
M.B.vanGijzen@TUDelft.nl 
Delft University of Technology, The Netherlands

Van Huffel, Sabine 
sabine.vanhuffel@esat.kuleuven.be 
Katholieke Universiteit Leuven, Belgium

Van Loan, Charles 
cv@cs.cornell.edu 
Cornell University, US

Vandebril, Raf 
raf.vandebril@cs.kuleuven.be 
Katholieke Universiteit Leuven, Belgium

Vandereycken, Bart 
bart.vandereycken@sam.math.ethz.ch 
ETH Zürich, Switzerland

Varah, Jim 
jmvarah@gmail.com 
The University of British Columbia, Canada

Varga, Richard 
varga@mcs.kent.edu 
Kent State University, US

Vavasis, Stephen A. 
vavasis@math.uwaterloo.ca 
University of Waterloo, Canada

Vecharynski, Eugene 
eugenev@cs.umn.edu 
University of Minnesota, US

Vu, Linh 
linhvh@vnu.edu.vn 
Vietnam National University, Vietnam

Vuik, Kees 
c.vuik@tudelft.nl 
Delft University of Technology, The Netherlands

Walker, Homer F. 
walker@wpi.edu 
Worcester Polytechnic Institute, US

Wathen, Andy 
wathen@maths.ox.ac.uk 
University of Oxford, UK

Watkins, David S. 
davidwatkins48@gmail.com 
Washington State University, US

Willems, Paul 
willems@math.uni-wuppertal.de 
WestLB AG, Germany

Xia, Jianlin 
xiaj@math.purdue.edu 
Purdue University, US

Xue, Fei 
fxue@temple.edu 
Temple University, US

Yamazaki, Ichitaro 
ic.yamazaki@gmail.com 
Lawrence Berkeley National Laboratory, US

Yang, Chao 
cyang@lbl.gov 
Lawrence Berkeley National Laboratory, US

Zhang, Bo 
zhangb@cs.duke.edu 
Duke University, US