

Matrix Methods in Data Mining and Pattern Recognition

Fundamentals of Algorithms

Editor-in-Chief: Nicholas J. Higham, University of Manchester

The SIAM series on Fundamentals of Algorithms is a collection of short user-oriented books on state-of-the-art numerical methods. Written by experts, the books provide readers with sufficient knowledge to choose an appropriate method for an application and to understand the method's strengths and limitations. The books cover a range of topics drawn from numerical analysis and scientific computing. The intended audiences are researchers and practitioners using the methods and upper level undergraduates in mathematics, engineering, and computational science.

Books in this series not only provide the mathematical background for a method or class of methods used in solving a specific problem but also explain how the method can be developed into an algorithm and translated into software. The books describe the range of applicability of a method and give guidance on troubleshooting solvers and interpreting results. The theory is presented at a level accessible to the practitioner. MATLAB® software is the preferred language for codes presented since it can be used across a wide variety of platforms and is an excellent environment for prototyping, testing, and problem solving.

The series is intended to provide guides to numerical algorithms that are readily accessible, contain practical advice not easily found elsewhere, and include understandable codes that implement the algorithms.

Editorial Board

Peter Benner
Technische Universität Chemnitz

John R. Gilbert
University of California, Santa Barbara

Michael T. Heath
University of Illinois, Urbana-Champaign

C. T. Kelley
North Carolina State University

Cleve Moler
The MathWorks

James G. Nagy
Emory University

Dianne P. O'Leary
University of Maryland

Robert D. Russell
Simon Fraser University

Robert D. Skeel
Purdue University

Danny Sorensen
Rice University

Andrew J. Wathen
Oxford University

Henry Wolkowicz
University of Waterloo

Series Volumes

Eldén, L., *Matrix Methods in Data Mining and Pattern Recognition*

Hansen, P. C., Nagy, J. G., and O'Leary, D. P., *Deblurring Images: Matrices, Spectra, and Filtering*

Davis, T. A., *Direct Methods for Sparse Linear Systems*

Kelley, C. T., *Solving Nonlinear Equations with Newton's Method*

Lars Eldén

Linköping University
Linköping, Sweden

Matrix Methods in Data Mining and Pattern Recognition

siam

Society for Industrial and Applied Mathematics
Philadelphia

Copyright © 2007 by the Society for Industrial and Applied Mathematics.

10 9 8 7 6 5 4 3 2 1

All rights reserved. Printed in the United States of America. No part of this book may be reproduced, stored, or transmitted in any manner without the written permission of the publisher. For information, write to the Society for Industrial and Applied Mathematics, 3600 University City Science Center, Philadelphia, PA 19104-2688.

Trademarked names may be used in this book without the inclusion of a trademark symbol. These names are used in an editorial context only; no infringement of trademark is intended.

Google is a trademark of Google, Inc.

MATLAB is a registered trademark of The MathWorks, Inc. For MATLAB product information, please contact The MathWorks, Inc., 3 Apple Hill Drive, Natick, MA 01760-2098 USA, 508-647-7000, Fax: 508-647-7101, info@mathworks.com, www.mathworks.com

Figures 6.2, 10.1, 10.7, 10.9, 10.11, 11.1, and 11.3 are from L. Eldén, Numerical linear algebra in data mining, *Acta Numer.*, 15:327–384, 2006. Reprinted with the permission of Cambridge University Press.

Figures 14.1, 14.3, and 14.4 were constructed by the author from images appearing in P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, Eigenfaces vs. fisherfaces: Recognition using class specific linear projection, *IEEE Trans. Pattern Anal. Mach. Intell.*, 19:711–720, 1997.

Library of Congress Cataloging-in-Publication Data

Eldén, Lars, 1944–

Matrix methods in data mining and pattern recognition / Lars Eldén.

p. cm. — (Fundamentals of algorithms ; 04)

Includes bibliographical references and index.

ISBN 978-0-898716-26-9 (pbk. : alk. paper)

1. Data mining. 2. Pattern recognition systems—Mathematical models. 3. Algebras, Linear. I. Title.

QA76.9.D343E52 2007

05.74—dc20

2006041348

Contents

Preface	ix
I Linear Algebra Concepts and Matrix Decompositions	
1 Vectors and Matrices in Data Mining and Pattern Recognition	3
1.1 Data Mining and Pattern Recognition	3
1.2 Vectors and Matrices	4
1.3 Purpose of the Book	7
1.4 Programming Environments	8
1.5 Floating Point Computations	8
1.6 Notation and Conventions	11
2 Vectors and Matrices	13
2.1 Matrix-Vector Multiplication	13
2.2 Matrix-Matrix Multiplication	15
2.3 Inner Product and Vector Norms	17
2.4 Matrix Norms	18
2.5 Linear Independence: Bases	20
2.6 The Rank of a Matrix	21
3 Linear Systems and Least Squares	23
3.1 LU Decomposition	23
3.2 Symmetric, Positive Definite Matrices	25
3.3 Perturbation Theory and Condition Number	26
3.4 Rounding Errors in Gaussian Elimination	27
3.5 Banded Matrices	29
3.6 The Least Squares Problem	31
4 Orthogonality	37
4.1 Orthogonal Vectors and Matrices	38
4.2 Elementary Orthogonal Matrices	40
4.3 Number of Floating Point Operations	45
4.4 Orthogonal Transformations in Floating Point Arithmetic . . .	46

5	QR Decomposition	47
5.1	Orthogonal Transformation to Triangular Form	47
5.2	Solving the Least Squares Problem	51
5.3	Computing or Not Computing Q	52
5.4	Flop Count for QR Factorization	53
5.5	Error in the Solution of the Least Squares Problem	53
5.6	Updating the Solution of a Least Squares Problem	54
6	Singular Value Decomposition	57
6.1	The Decomposition	57
6.2	Fundamental Subspaces	61
6.3	Matrix Approximation	63
6.4	Principal Component Analysis	66
6.5	Solving Least Squares Problems	66
6.6	Condition Number and Perturbation Theory for the Least Squares Problem	69
6.7	Rank-Deficient and Underdetermined Systems	70
6.8	Computing the SVD	72
6.9	Complete Orthogonal Decomposition	72
7	Reduced-Rank Least Squares Models	75
7.1	Truncated SVD: Principal Component Regression	77
7.2	A Krylov Subspace Method	80
8	Tensor Decomposition	91
8.1	Introduction	91
8.2	Basic Tensor Concepts	92
8.3	A Tensor SVD	94
8.4	Approximating a Tensor by HOSVD	96
9	Clustering and Nonnegative Matrix Factorization	101
9.1	The k -Means Algorithm	102
9.2	Nonnegative Matrix Factorization	106
II Data Mining Applications		
10	Classification of Handwritten Digits	113
10.1	Handwritten Digits and a Simple Algorithm	113
10.2	Classification Using SVD Bases	115
10.3	Tangent Distance	122
11	Text Mining	129
11.1	Preprocessing the Documents and Queries	130
11.2	The Vector Space Model	131
11.3	Latent Semantic Indexing	135
11.4	Clustering	139

11.5	Nonnegative Matrix Factorization	141
11.6	LGK Bidiagonalization	142
11.7	Average Performance	145
12	Page Ranking for a Web Search Engine	147
12.1	Pagerank	147
12.2	Random Walk and Markov Chains	150
12.3	The Power Method for Pagerank Computation	154
12.4	HITS	159
13	Automatic Key Word and Key Sentence Extraction	161
13.1	Saliency Score	161
13.2	Key Sentence Extraction from a Rank- k Approximation	165
14	Face Recognition Using Tensor SVD	169
14.1	Tensor Representation	169
14.2	Face Recognition	172
14.3	Face Recognition with HOSVD Compression	175
 III Computing the Matrix Decompositions		
15	Computing Eigenvalues and Singular Values	179
15.1	Perturbation Theory	180
15.2	The Power Method and Inverse Iteration	185
15.3	Similarity Reduction to Tridiagonal Form	187
15.4	The QR Algorithm for a Symmetric Tridiagonal Matrix	189
15.5	Computing the SVD	196
15.6	The Nonsymmetric Eigenvalue Problem	197
15.7	Sparse Matrices	198
15.8	The Arnoldi and Lanczos Methods	200
15.9	Software	207
Bibliography		209
Index		217

Preface

The first version of this book was a set of lecture notes for a graduate course on data mining and applications in science and technology organized by the Swedish National Graduate School in Scientific Computing (NGSSC). Since then the material has been used and further developed for an undergraduate course on numerical algorithms for data mining and IT at Linköping University. This is a second course in scientific computing for computer science students.

The book is intended primarily for undergraduate students who have previously taken an introductory scientific computing/numerical analysis course. It may also be useful for early graduate students in various data mining and pattern recognition areas who need an introduction to linear algebra techniques.

The purpose of the book is to demonstrate that there are several very powerful numerical linear algebra techniques for solving problems in different areas of data mining and pattern recognition. To achieve this goal, it is necessary to present material that goes beyond what is normally covered in a first course in scientific computing (numerical analysis) at a Swedish university. On the other hand, since the book is application oriented, it is not possible to give a comprehensive treatment of the mathematical and numerical aspects of the linear algebra algorithms used.

The book has three parts. After a short introduction to a couple of areas of data mining and pattern recognition, linear algebra concepts and matrix decompositions are presented. I hope that this is enough for the student to use matrix decompositions in problem-solving environments such as MATLAB®. Some mathematical proofs are given, but the emphasis is on the existence and properties of the matrix decompositions rather than on how they are computed. In Part II, the linear algebra techniques are applied to data mining problems. Naturally, the data mining and pattern recognition repertoire is quite limited: I have chosen problem areas that are well suited for linear algebra techniques. In order to use intelligently the powerful software for computing matrix decompositions available in MATLAB, etc., some understanding of the underlying algorithms is necessary. A very short introduction to eigenvalue and singular value algorithms is given in Part III.

I have not had the ambition to write a book of recipes: “given a certain problem, here is an algorithm for its solution.” That would be difficult, as the area is far too diverse to give clear-cut and simple solutions. Instead, my intention has been to give the student a set of tools that may be tried as they are but, more likely, that will need to be modified to be useful for a particular application. Some of the methods in the book are described using MATLAB scripts. They should not

be considered as serious algorithms but rather as pseudocodes given for illustration purposes.

A collection of exercises and computer assignments are available at the book's Web page: www.siam.org/books/fa04.

The support from NGSSC for producing the original lecture notes is gratefully acknowledged. The lecture notes have been used by a couple of colleagues. Thanks are due to Gene Golub and Saara Hyvönen for helpful comments. Several of my own students have helped me to improve the presentation by pointing out inconsistencies and asking questions. I am indebted to Berkant Savas for letting me use results from his master's thesis in Chapter 10. Three anonymous referees read earlier versions of the book and made suggestions for improvements. Finally, I would like to thank Nick Higham, series editor at SIAM, for carefully reading the manuscript. His thoughtful advice helped me improve the contents and the presentation considerably.

Lars Eldén
Linköping, October 2006