APPLICATIONS OF LINEAR ALGEBRA

TO DNA MICROARRAYS

Amir Niknejad Shmuel Friedland

In memory of our parents: Asghar Niknejad, Aron and Golda Friedland

Preface

Linear Algebra has been used in many scientific and engineering disciplines: computer science, as computer graphics; physics, as quantum mechanics; electrical engineering, as signal processing; mechanical engineering, as robotics. Most recently, Linearly Algebra emerged in life science as in computational molecular biology.

The main motivation of this monograph is condensing the information which arises in molecular biology in general, and in gene expression data imbedded in DNA microarrays in particular. The second motivation is to complete the missing data in the above instances. Our main tools are coming from the linear algebra, in particular approximation of matrices by low rank matrices, as given by *Singular Value Decomposition*, abbreviated as SVD, or other appropriate matrix decompositions.

In the last few years it became clear that a more general way to approach multidimensional in biological data, as well as an engineering data, is to use tensors analysis, rather than matrix analysis. The last Chapter of this monograph deals with basic aspects of tensors, as low rank approximations, which is the analog of low rank approximations of matrices discussed in details in this monograph. Most of the material in this monograph is based on the Ph.D. thesis of the first named author [36].

We now summarize briefly the contents of this monograph. In Chapter 1 we give first some biological background for gene expressions and DNA microarrays. Next we introduce the singular value decomposition (SVD) of matrices and its extensions. We then mention some applications of SVD in analyzing gene expression data, image processing and information retrieval.

In Chapter 2 we introduce randomized low rank approximations of matrices which do not use SVD. We present our Monte Carlo algorithm to achieve such an approximation, along with other algorithms.

Chapter 3 deals with clustering methods and their application in the analysis of the gene expression data. We introduce several clustering methods such as the ε-clustering, K-means clustering, spectral clustering and EM clustering algorithm.

Chapter 4 deals with imputing missing data in gene expression of micro-arrays. We introduce our fixed rank approximation algorithm (FRAA) for imputing missing data in the DNA gene expression array. Finally, we use simulation to compare FRAA versus other methods and indicate the advantages and its shortcomings, and how to overcome the shortcomings of FRAA.

Chapter 5 deals with basic concepts of tensors, which are related to the topics discussed in Chapters 1 - 4, mostly low rank approximations of tensors.

Amir Niknejad, College of Mount Saint Vincent,
Shmuel Friedland, University of Illinois at Chicago.

Contents

1 Introduction and Background **1**

1.1 Missing gene imputation in DNA microarrays 1

1.2 Some biological background 1

1.2.1 Transcription and Translation 2

1.2.2 DNA microarrays (chips) 2

1.3 Some linear algebra background 3

1.3.1 Inner Product Spaces (IPS) 4

1.3.2 Definition and Properties of Positive Definite and Semidefinite Matrices 4

1.3.3 Various Matrix Factorizations 5

1.3.4 Gram-Schmidt Process 5

1.3.5 Cholesky Decomposition 6

1.3.6 The QR Decomposition 6

1.3.7 An Introduction to the Singular Value Decomposition 7

1.3.8 SVD on inner product spaces 9

1.4 Extended Singular Value decomposition 10

1.5 Some Applications of SVD 13

1.5.1 Analysing DNA gene expression data via SVD 13

1.5.2 Low rank approximation of matrices 13

2 Randomized Low Rank Approximations **14**

2.1 Random Projection Method 15

2.2 Fast Monte-Carlo Rank Approximation 17

2.3 CUR approximation 21

3 Cluster Analysis **23**

3.1 Clusters and Clustering 23

3.2 Various Gene Expression Clustering Procedures 23

3.2.1 Proximity (Similarity) Measurement 23

3.2.2 ε-clustering 24

3.2.3 Convex ε-clustering 24

3.2.4 Hierarchical Cluster Analysis 25

3.2.5 Clustering Using K-means Algorithm 25

3.2.6 Spectral clustering 26

3.2.7 Mixture Models and EM Algorithm 27

4 Various Methods of Imputation **28**

4.1 The Gene Expression Matrix 28

4.2 SVD and gene clusters 29

4.3 Missing Data in the Gene Expression Matrix 30
4.3.1 Reconsideration of 4.2 30
4.3.2 Bayesian Principal Component Analysis (BPCA) 31
4.3.3 The least square imputation method 33
4.3.4 Iterative method using SVD 34
4.3.5 Local least squares imputation (LLSimpute) 34
4.4 Motivation for FRAA 35
4.4.1 Additional matrix theory background for FRAA 35
4.4.2 The Optimization Problem 36
4.5 Fixed Rank Approximation Algorithm 37
4.5.1 Description of FRAA 37
4.5.2 Explanation and justification of FRAA 38
4.5.3 Algorithm for (4.14) 41
4.6 Simulation 41
4.7 Discussion of FRAA 44
4.8 IFRAA 44
4.8.1 Introduction 44
4.8.2 Computational comparisons of
BPCA, FRAA, IFRAA and LLSimpute 45
4.9 Conclusions 49
4.10 Matlab code 50

5 Tensors 52

5.1 Motivation for tensors 52
5.2 Basic notions and results 52
5.3 Inner products on tensor spaces 53
5.4 The SVD as best subspace tensor approximation 55
5.5 Best subspace tensor approximations for 3-tensors 55
5.6 Fast low rank 3-tensors approximations 57

Bibliography 58

Notations 62

1 Introduction and Background

1.1 Missing gene imputation in DNA microarrays

1.2 Some biological background

1.2.1 Transcription and Translation

1.2.2 DNA microarrays (chips)

1.3 Some linear algebra background

1.3.1 Inner Product Spaces (IPS)

1.3.2 Definition and Properties of Positive Definite and Semidefinite Matrices

1.3.3 Various Matrix Factorizations

1.3.4 Gram-Schmidt Process

1.3.5 Cholesky Decomposition

1.3.6 The QR Decomposition

1.3.7 An Introduction to the Singular Value Decomposition

1.3.8 SVD on inner product spaces

1.4 Extended Singular Value decomposition

1.5 Some Applications of SVD

1.5.1 Analysing DNA gene expression data via SVD

1.5.2 Low rank approximation of matrices

2 Randomized Low Rank Approximations of Matrices

2.1 Random Projection Method

Low rank approximation (FKV method).

2.2 Fast Monte-Carlo Rank Approximation

2.3 CUR approximation

3 Cluster Analysis

3.1 Clusters and Clustering

3.2 Various Gene Expression Clustering Procedures

3.2.1 Proximity (Similarity) Measurement

3.2.2 ε-clustering

3.2.3 Convex ε-clustering

3.2.4 Hierarchical Cluster Analysis

One possibility for

3.2.5 Clustering Using K-means Algorithm

3.2.6 Spectral clustering

3.2.7 Mixture Models and EM Algorithm

4 Various Methods of Imputation of Missing Values in DNA Microarrays

4.1 The Gene Expression Matrix

4.2 SVD and gene clusters

4.3 Missing Data in the Gene Expression Matrix

4.3.1 Reconsideration of 4.2

4.3.2 Bayesian Principal Component Analysis (BPCA)

4.3.3 The least square imputation method

4.3.4 Iterative method using SVD

4.3.5 Local least squares imputation (LLSimpute)

4.4 Motivation for FRAA

4.4.1 Additional matrix theory background for FRAA

4.4.2 The Optimization Problem

4.5 Fixed Rank Approximation Algorithm

4.5.1 Description of FRAA

4.5.2 Explanation and justification of FRAA

4.5.3 Algorithm for (4.14)

4.6 Simulation

4.7 Discussion of FRAA

4.8 IFRAA

4.8.1 Introduction

4.8.2 Computational comparisons of BPCA, FRAA, IFRAA and LLSimpute

4.9 Conclusions

4.10 Matlab code

5 Tensors

5.1 Motivation for tensors

5.2 Basic notions and results

5.3 Inner products on tensor spaces

5.4 The SVD as best subspace tensor approximation

5.5 Best subspace tensor approximations for 3-tensors

5.6 Fast low rank 3-tensors approximations

References

[1] C.C. Aggrawal, C.M. Procopiuc, J.L. Wolf, P.S. Yu and J.S. Park, Fast algorithms for projected clustering, *Proc. of ACM SIGMOD Intl. Conf. Management of Data* 1999, 61-72.

[2] R. Agrawal, J.Gerhrke, D.Gunopulos, and P. Raghavan, Automatic subspace clustering of high dimensional data for data mining applications, *Proc. ACM SIGMOD Conf. on Management of Data*, 1998, 94-105.

[3] U. Alon et al., Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc. Natl. Acad. Sci. USA* 96 (1999), 6745-6750.

[4] O. Alter, P.O. Brown and D. Botstein, Processing and modelling gene expression expression data using the singular value decomposition, *Proceedings SPIE*, vol. 4266 (2001), 171-186.

[5] O. Alter, P.O. Brown and D. Botstein, Generalized singular decomposition for comparative analysis of genome-scale expression data sets of two different organisms, *Proc. Nat. Acad. Sci. USA* 100 (2003), 3351-3356.

[6] O. Alter, G.H. Golub, P.O. Brown and D. Botstein, Novel genome-scale correlation between DNA replication and RNA transcription during the cell cycle in yeast is predicted by data-driven models, 2004 *Miami Nature Winter Symposium*, Jan. 31 - Feb. 4, 2004.

[7] P. Baldi and G. Wesley Hatfield, *DNA Microarrays and Gene Expression*, Cambridge University Press, 2002.

[8] M. Belkin and P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, *Neural Computation* 15 (2003), 1373-1396.

[9] T.H. Bo, B. Dysvik and I. Jonassen, LSimpute, Accurate estimation of missing values in microarray data with least squares methods, *Nucleic Acids Research*, 32 (2004), e34.

[10] H. Chipman, T.J. Hastie and R. Tibshirani, Clustering micrarray data In: T. Speed, (Ed.), *Statistical Analysis of Gene Expression Microarray Data*, Chapman & Hall/CRC, 2003 pp. 159-200.

[11] E. Domany, Cluster Analysis of Gene Expression Data, *Journal of Statistical Physics*, 110 (2003), 1117-1139

[12] P. Drineas, A. Frieze, R. Kannan, S. Vempala and V. Vinay, Clustering large graphs via the singular value decomposition, *Journal of Machine Learning*, 56 (2004), 9-33.

[13] Petros Drineas, Eleni Drinea, Patrick S. Huggins,*An Experimental Evaluation of a Monte-Carlo Algorithm for Singular Value Decomposition*, Panhellenic Conference on Informatics 2001: 279-296

[14] M. Ester, H.-P. Krieger, J. Sander and X.Xu, A density-based algortihm for discovering clusters in large spatial databases with nose, *Proc. 2nd Intl. Conf. Knowledge Discovery and Data Mining*, 1996, 226-231.

[15] S. Friedland, Inverse eigenvalue problems, *Linear Algebra* Appl., 17 (1977), 15-51.

[16] S. Friedland, On the generic rank of 3-tensors, arXiv:0805.3777v2.

[17] S. Friedland, M. Kaveh, A. Niknejad, H. Zare, An algorithm for missing value estimation for DNA microarray data, (with M. Kaveh, A. Niknejad, H. Zare), IEEE *Proceedings of ICASSP 2006*, II, 1092-10095.

[18] S. Friedland, M. Kaveh, A. Niknejad, H. Zare, Fast Monte-Carlo low rank approximations for matrices, *Proc. IEEE Conference SoSE*, Los Angeles, 2006, 218-223.

[19] S. Friedland and V. Mehrmann, Best subspace tensor approximations, arXiv:0805.4220v1

[20] S. Friedland, V. Mehrmann, A. Miedlar and M. Nkengla Fast low rank approximations of matrices and tensors, submitted, http://www.matheon.de/research/show_preprint.asp?action=details&serial=456.

[21] S. Friedland, A. Niknejad and L. Chihara, A Simultaneous Reconstruction of Missing Data in DNA Microarrays, *Linear Alg. Appl.*, 416 (2006), 8-28.

[22] S. Friedland, J. Nocedal and M. Overton, The formulation and analysis of numerical methods for inverse eigenvalue problems, *SIAM J. Numer. Anal.* 24 (1987), 634-667.

[23] S. Friedland and A. Torokhti. Generalized rank-constrained matrix approximations, *SIAM J. Matrix Anal. Appl.* 29 (2007), 656 – 659.

[24] A. Frieze, R. Kannan and S. Vempala, Fast Monte-Carlo alogrithms for finding low rank approximations, *Proceedings of the 39th Annual Symposium on Foundation of Computer Science*, 1998.

[25] D. Fritzsche, V. Mehrmann, D. Szyld, E. Virnik, An SVD approach to identifying metastable states of Markov chains *Electronic Transactions on Numerical Analysis*, 29 (2008), 46-89.

[26] X. Gan, A.W.-C. Liew and H. Yan, Missing Microaaray Data Estimation Based on Projection onto Convex Sets Method, *Proc. 17th International Conference on Pattern Recognition*, 2004.

[27] G.H. Golub and C.F. Van Loan, *Matrix Computations*, John Hopkins Univ. Press, 1983.

[28] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science* 286 (1999), 531-537.

[29] S.A. Goreinov, E.E. Tyrtyshnikov, and N.L. Zamarashkin, A theory of pseudo-skeleton approximations of matrices, *Linear Algebra Appl.* 261 (1997), 1 – 21.

[30] R.A. Horn and C.R. Johnson, *Matrix analysis*, Cambridge Univ. Press, 1987.

[31] D.A. Jackson, Stopping rules in principal component analysis: a comparison of heuristical and statistical approaches, *Ecology* 74 (1993), 2204-2214.

[32] D. Jiang , C. Tang and A.Zhang, Cluster Analysis for Gene Expression Data: A Survey, *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, Volome 16(11), page 1370 - 1386, 2004

[33] R.A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis, Prentice Hall*, New Jersey, 4th edition (1998).

[34] H. Kim, G.H. Golub and H. Park, Missing value estimation for DNA microarray gene expression data: local least squares imputation, *Bioinformatics* 21 (2005), 187-198.

[35] A.Y. Ng, M.I. Jordan and Y. Weiss, On spectral clustering: Analysis and an algorithm, in T.K. Leen, T.G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems* 14, MIT Press, 2002.

[36] A. Niknejad, *Application of Singular Value Decomposition to DNA Microarray*, Ph.D. thesis, University of Illinois at Chicago, 2005.

[37] S. Oba, M. Sato, I. Takemasa, M. Monden, K. Matsubara and S. Ishii, A Baesian missing value estimation method for gene expression profile data, *Bioinformatics* 19 (2003), 2088-2096.

[38] C.C. Paige and M. A. Saunders, Towards a generalized singular value decomposition, *SIAM J. Numer. Anal.* 18 (1981), 398-405.

[39] C.M. Procopiuc, P.K. Agarwal, M. Jones and T.M. Murali, A Monte Carlo algorithm for fast projective clustering, *Proc. of ACM SIGMOD Intl. Conf. Management of Data* 2002.

[40] M.A. Shipp, K.N. Ross, P. Tamayo, A.P. Weng, J.L. Kutok, R.C. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G.S. Pinkus *et al.*, Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling nad supervised machine learning, *Nat. Med.* 8 (2002), 68-74.

[41] A. Schulze and J. Downward, *Nature Cell. Biol.* 3:190 (2001)

[42] R. Shioda and L. Tuncel, Clustering via minimum volume ellipsoids, *Comput. Optim. Appl.* 37 (2007), 247-295.

[43] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein and B. Futcher, Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, *Mol. Biol. Cell*, **9** (1998), 3273-3297.

[44] G.W. Stewart, A method for computing the generalized singular value decomposition, *Matrix Pencils*, B. Kagström and A. Ruhe, *Lecture Notes in Mathematics*, 973 (1982), 207-220.

[45] M. Tamura, Missing value expectation of matrix data by fixed rank approximation algorithm, Master Project Report, Computer Science Department, University of Illinois at Chicago, May 2006, http://www.cs.uic.edu/ mtamura/MakioTamuraMasterProject.pdf

[46] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R. Altman, Missing value estimation for DNA microarrays, *Bioinformatics* 17 (2001), 520-525.

[47] L.R. Tucker. Some mathematical notes on three-mode factor analysis, *Psychometrika* 31 (1966), 279 – 311.

[48] S. Vempala, The Random Projection Method, *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, vol. 65, American Mathematical Society, 2004.

Notations

$\mathbb{F}$: the field of real numbers $\mathbb{R}$ or the field of complex numbers $\mathbb{C}$.
$\mathbb{F}^{m\times n}$: the space of all $m \times n$ matrices with entries in $\mathbb{F}$.
$S_m(\mathbb{R})$: the set of $n \times n$ real symmetric matrices.
I_m: the $m \times m$ identity matrix.
$\mathrm{O}_{mk}(\mathbb{R})$: the set of $m \times k$ real matrices O satisfying $O^{\mathrm{T}}O = I_k$.
$\mathbb{F}^{m\times n\times l}$: the space of all $m \times n \times l$ tensors with entries in $\mathbb{F}$.
$\mathbf{U}$: or other bold capital letter denotes a vector space.
$\mathbf{u}$: or other bold lower case denotes a vector space.
a: or other lower case may denote scalar.
α: or other lower case Greek letter may denote scalar.
A: or other capital letter denotes matrix, with the entries a_{ij}.
$\mathcal{A}$: or other calligraphic capital letter denotes 3-tensor, with the entries a_{ijk}.
$\otimes$: tensor product.
$\langle \mathbf{x}, \mathbf{y}\rangle$: inner product of the vectors $\mathbf{x}, \mathbf{y}$.
$||\mathbf{x}||$: norm of $\mathbf{x}$, usually Hilbert norm $||\mathbf{x}|| = \sqrt{\langle \mathbf{x}, \mathbf{x}\rangle}$.
$||A||_{\mathcal{F}}$: The Frobenius norm of A, where A viewed as a vector with mn coordinates.
$\mathbf{x}^{\mathrm{T}}, A^{\mathrm{T}}$: the transpose of a vector and a matrix, respectively.
$A > 0, A \geq 0$: a symmetric positive definite matrix, nonnegative definite matrix, respectively.
$A[I, J], A_{I,J}, A_{IJ}$: the submatrix of A with the rows and columns in the sets I, J respectively.
$\mathcal{A}_{IJK}$: the 3-subtensor of $\mathcal{A}$ with the indices in the sets I, J, K respectively.
$\mathcal{R}(E), \mathcal{N}(E)$: the range and the null space of E respectively.
$\mathrm{span}(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ the subspace spanned by $\mathbf{x}_1, \ldots, \mathbf{x}_k$.
$\mathrm{rank}\, A$: the rank of A.
$\sigma_i(A)$: the $i - th$ singular value of A.
$\lambda_i(A)$: the $i - th$ eigenvalue of symmetric A.
$\mathrm{diag}(d_1, \ldots, d_m)$: diagonal matrix with $d_1, \ldots, d_m$ on the diagonal.
T^*: is $\overline{T}^{\mathrm{T}}$ for a matrix T, and the adjoint operator for an operator T.
$\mathcal{R}(n, m, k)$: the set of all $m \times n$ matrices of rank k at most.
$E^{\dagger}$: Moore-Penrose generalized inverse of E.
$\langle m\rangle$: the set of integers from 1 to m, where m is a positive integer.
$\mathrm{Gr}(q, \mathbb{F}^m)$: The set of all q-dimensional subspaces in $\mathbb{F}^m$.
KNN: K-nearest neighbor clustering algorithm.
IPS: abbreviation for *Inner Product space*
SVD: abbreviation for *Singular Value Decomposition.*
ESVD: *Extended Singular Value Decomposition*
BPCA: Bayesian Principal Component Analysis.
FRAA: Fixed Rank Approximation Algorithm.
IFRAA: Improved Fixed Rank Approximation Algorithm.

LLSimpute: Local Least Squares imputation.
NRMSE: Normalized Root Mean Square Error.
RMSE: Root Mean Square Error.