The Role of Singular Value Decomposition in Gene Expression Microarrays

Amir Niknejad
University of Illinois at Chicago

UVW, April 3, 2006

1 Outline of the talk

1. Estimation of missing values in given matrix data using the inverse eigenvalue problems techniques, and their applications to DNA microarrays and image processing.

2. A joint SVD decomposition of two or more matrices to compare several biological processes.

Most of the results can be found in the following recent papers, which are available at

http://www.math.uic.edu/∼friedlan/research.html

Singular Value Decomposition

$G^{(n\times m)} = U^{(n\times n)}\Sigma^{(m\times m)}V^{(m\times n)^T}, n \gg m$

$\Sigma = \text{DIAG}(\sigma_1, \sigma_2, \ldots, \sigma_m)$

$\sigma_i = \sqrt{\lambda_i}$ **eigenvalues of** GG^T **or** G^TG

Columns of U**, eigenvectors of** GG^T**.**

Columns of V**, eigenvectors of** G^TG**.**

Let $\text{rank}(G) = r$**. Then**

- $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \sigma_r > \sigma_{r+1} = \cdots \sigma_m = 0$
- $N(G) = \text{span}\{v_{r+1}, \ldots, v_n\}$ **and** $R(G) = \text{span}\{u_1, \ldots, u_r\}$
- $G = \Sigma_{i=1}^{r} u_i v_i v_i^T$
- **Norms:** $\|G\|_F^2 = \sigma_1^2 + \cdots + \sigma_r^2$**, and** $\|G\|_2^2 = \sigma_1$

Principal Component Analysis (PCA)

- **Let $G = (\mathrm{g}_1, \ldots, \mathrm{g}_n)^T$, genes through different experiments. Let $\mu_G = E(G)$ be the mean and $C_G = (G - \mu_G)(G - \mu_G)^T$ be the covariance matrix.**
- **Then compute the eigenvalues λ_i and corresponding eigenvector $v_i, i = 1 \ldots n$.**
- **Sort the eigenvalues and corresponding eigenvectors in descending order.**
- **Construct orthogonal basis, by having the first eigenvector the direction of the largest variance.**
- **Represent the data with the few basis vector (corresponding to the largest eigen value).**
- **Let E_k be the matrix having the first k eigenvectors.**
- **Project the original data on the axis in dimension k. This way we could minimize MSE, by this representation of data. This way we lose a little bit of information.**

Low Rank Approximation

Let $\mathcal{R}(n, m, k)$ denote the set of $n \times m$ matrices of at most rank k ($m \geq k$). Then for each k, $k \leq r$, the SVD of E gives the solution to the following approximation problem:

$$\min_{E \in \mathcal{R}(n,m,k)} ||G - E||_{\mathcal{F}} ||G - \sum_{q=1}^{k} \sigma_q \mathbf{u}_q \mathbf{v}_q^{\mathrm{T}}||_{\mathcal{F}}$$

$$= \sqrt{\sum_{q=k+1}^{r} \sigma_q^2}$$

Transcriptional Response:

$$\mathbf{g}_i = \sum_{k=1}^{r} u_{ik} \sigma_k \mathbf{v}_k$$

Array expression profile:

$$\mathbf{c}_j = \sum_{k=1}^{r} \mathbf{v}_{jk} \sigma_k \mathbf{u}_k$$

Computing Eigen-genes

Gene matrix $G^{n \times m}$, # of significant $\sigma_i \leqslant \frac{m}{2}$

$\mathbf{u}_1, \ldots, \mathbf{u}_n$ eigenarrays, $\mathbf{v}_1, \ldots, \mathbf{v}_n$ eigengenes, $\sigma_1 \ldots, \sigma_l$ eigenexpressions.

- **Hous**$(G) = (I - 2\mathbf{h}\mathbf{h}^{\mathbf{t}})\mathbf{G} = \mathbf{A}.\|\mathbf{h}\|_2 = 1$
- **Compute** $A^T A$
- $QR(A^T A) \rightarrow Q \begin{bmatrix} R \\ 0 \end{bmatrix} \rightarrow$

  positive $\sigma_1^2, \ldots, \sigma_l^2$ **and** $\mathbf{u}_1, \ldots, \mathbf{u}_n$

2 SVD in inner product spaces

$\mathbf{U}_i$ is m_i-dimensional IPS over $\mathbb{C}$, with $\langle \cdot, \cdot \rangle_i, i = 1, 2$.
$T : \mathbf{U}_1 \to \mathbf{U}_2$ linear operator. $T^* : \mathbf{U}_2 \to \mathbf{U}_1$ the adjoint operator: $\langle T\mathbf{x}, \mathbf{y} \rangle_2 = \langle \mathbf{x}, T^*\mathbf{y} \rangle_1$.
$S_1 := T^*T : \mathbf{U}_1 \to \mathbf{U}_1$,
$S_2 := TT^* : \mathbf{U}_2 \to \mathbf{U}_2$.

S_1, S_2 self-adjoint: $S_1^* = S_1, S_2^* = S_2$ and nonnegative definite: $\langle S_i\mathbf{x}_i, \mathbf{x}_i \rangle_i \geq 0$.

$\sigma_1^2 \geq ... \geq \sigma_r^2 > 0$ positive eigenvalues of S_1 and S_2 and $r = \operatorname{rank} T = \operatorname{rank} T^*$. Let
$S_1\mathbf{v}_i = \sigma_i^2\mathbf{v}_i, \ \langle \mathbf{v}_i, \mathbf{v}_j \rangle_1 = \delta_{ij}, \ i, j =, 1, ..., r.$
Define $\mathbf{u}_i := \sigma_i^{-1}T\mathbf{v}_i, i = 1, ..., r$. Then
$\langle \mathbf{u}_i, \mathbf{u}_j \rangle_2 = \delta_{ij}, i, j = 1, \ldots, r.$

Complete $\{\mathbf{v}_1, ..., \mathbf{v}_r\}$ and $\{\mathbf{u}_1, ..., \mathbf{u}_r\}$ to orthonormal bases $[\mathbf{v}_1, ..., \mathbf{v}_{m_1}]$ and $[\mathbf{u}_1, ..., \mathbf{u}_{m_2}]$ in $\mathbf{U}_1$ and $\mathbf{U}_2$.

3 Matrix SVD

Let $A \in \mathbb{C}^{m\times n}$. Then $A : \mathbb{C}^n \longrightarrow \mathbb{C}^m$. Assume $\mathbb{C}^n, \mathbb{C}^m$ equipped with standard inner product $\langle \mathbf{x}, \mathbf{y} \rangle := \mathbf{y}^*\mathbf{x}$.

Then $A = U\Sigma V^*$, where $U \in \mathbf{U}(m), V \in \mathbf{U}(n)$, $\Sigma = \mathbf{diag}(\sigma_1, \ldots, \sigma_{\mathbf{min}(m,n)}) \in \mathbb{R}_+^{m\times n}$.

U, V transition matrices from $[\mathbf{u_1}, ..., \mathbf{u}_m], [\mathbf{v_1}, ..., \mathbf{v}_n]$ to the standard bases in $\mathbb{C}^m, \mathbb{C}^n$ respectively.

For $k \leq r$ let $\Sigma_k = diag(\sigma_1, \ldots, \sigma_k) \in \mathbb{R}^{k\times k}$, and $U_k \in \mathbf{U}(m, k), V_k \in \mathbf{U}(n, k)$ having the first k columns of U, V respectively. Then $A_k := U_k\Sigma_k V_k^*$ the best rank k approximation in Frobenius and operator norm of A:

$\mathbf{min}_{B\in\mathcal{R}(m,n,k)} ||A - B|| = ||A - A_k||$.

$A = U_r\Sigma_r V_r^*$ is Reduced SVD

$(r \geq)\ \nu$ numerical rank of A if $\frac{\sigma_{\nu+1}}{\sigma_\nu} \approx \mathbf{0}$.

A_ν is a noise reduction of A.

Noise reduction has many applications in image processing, DNA-Microarrays analysis, data compression.

4 MISSING ENTRIES PROBLEM

In DNA Microarrays experiments one measure thousands of genes $i = 1, \ldots, m$ in n different conditions, typically $n \in [3, 20]$.

$0 \leq a_{ij}$ measures the intensity of gene i in $j - th$ experiment. The results are recorded in the matrix $A = (a_{ij}) \in \mathbb{R}^{m \times n}$.

Sometimes the entries a_{ij} are missing (corrupted, up to 20%).

Let $\mathcal{T} \subset \{1, \ldots, n\} \times \{1, \ldots, m\}$ missing entries set.

Set $a_{ij} = 0$ if $(i, j) \in \mathcal{T}$.

Let $\mathcal{X}$ be all $X = (x_{ij}) \in \mathbb{R}^{m \times n}$ where $x_{ij} = 0$ if $(i, j) \notin \mathcal{T}$.

Assume that the completed matrix of the experiment should have the numerical rank ν. Then we complete the entries by solving the problem:

$(1) \quad \min_{X \in \mathcal{X}} \sum_{i=\nu+1}^{n} \sigma_i^2(A + X) = \min_{X \in \mathcal{X}} \sum_{i=\nu+1}^{n} \lambda_i((A + X)^{\mathrm{T}}(A + X))$

5 FRAA

Fixed Rank Approximation Algorithm: [8]

Let $G_p \in \mathcal{X}$ be the p^{th} approximation to a solution of optimization problem (1). Let $B_p := (A + G_p)^{\mathrm{T}}(A + G_p)$ and find an orthonormal set of eigenvectors for B_p, $\mathbf{v}_{p,1}, \ldots, \mathbf{v}_{p,m}$. Then G_{p+1} is a solution to the following minimum of a convex nonnegative quadratic function $\min_{X \in \mathcal{X}} \sum_{q=l+1}^{m} ((A + X)\mathbf{v}_{p,q})^{\mathrm{T}}((A + X)\mathbf{v}_{p,q})$.

Flow chart of the algorithm:

Fixed Rank Approximation Algorithm (FRAA)

Input: integers $m, n, L, iter$, the locations of non-missing entries $\mathcal{S}$, initial approximation G_0 of $n \times m$ matrix G.

Output: an approximation G_{iter} of G.

for $p = 0$ **to** $iter - 1$

- Compute $B_p := (A + G_p)^{\mathrm{T}}(A + G_p)$ and find an orthonormal set of eigenvectors for B_p, $\mathbf{v}_{p,1}, \ldots, \mathbf{v}_{p,m}$.
- G_{p+1} is a solution to the minimum problem (1) with $\nu = L - 1 = l$.

Let $f_l(X) := \sum_{i=\nu+1}^{n} \sigma_i^2(A+X)$. In each step of the algorithm $f_l(G_p) \geq f_l(G_{p+1})$. $G_p, p = 1, \ldots$ converges to a critical point $\tilde{G}$. FRAA gives a good approximation of $\tilde{G}$. In many simulations $\tilde{G} = G^*$.

FRAA is an adaptation of an algo for IEP:

Inverse Eigenvalue Problem: *Find the values of the missing entries of* G *such that the nonnegative definite matrix* $G^{\mathrm{T}}G$ will have $m - l$ smallest eigenvalues equal to zero.

IEP appear often in engineering. See [9] for examples of IEP and a number of good algorithms to solve these problems.

FRAA is a robust algorithm which performs good, but not as well as KNNimpute, BPCA and LSSimpute.

All other algo reconstruct the missing values of each gene from similar genes.

More about FRAA

Proposition

Let G' be the $n' \times m'$ be the matrix constructed by deleting i and j columns, the $\sigma_q(G) \geq \sigma_q(G')$ for $q = 1 \ldots m$.

Ky-Fan Characterization

$$\sum_{q=l}^{m} \lambda_q(A) = \sum_{q=l}^{m} \mathbf{z}_q^T A \mathbf{z}_q$$

$$= \min_{\{\mathbf{y}_l, \ldots, \mathbf{y}_m\} \in \Omega_{m-l+1}} \sum_{q=l}^{m} \mathbf{y}_q^T A \mathbf{y}_q$$

Computational Aspect of (FRAA)

The algorithm tries to minimize $\sum_{l=L}^{M} \sigma_{l+1}^{2}(G)$, where G ranges over all the completion of the missing data. We assume that the ideal completed gene matrix should have rank (k).

- We start with G_0 using SVD. Suppose we have G_p . G_p is the missing reconstruction data.
- Then algorithm for constructing G_{p+1},

$$A_p := G_p^T G_p \text{ an } n \times m \text{ matrix.}$$

- We use MatLab to find

$$\boldsymbol{\lambda_1(G_p) \geq \cdots \geq \lambda_m(G_p),}$$

and their corresponding eigenvector $\mathbf{u_{p,1}} \ldots \boldsymbol{u_{p,m}}$.

$\boldsymbol{G = G_p + X}$.

$\boldsymbol{X}$ has all zero entries for the non missing data. Now we consider the quadratic form,

$$\boldsymbol{\sum_{i=l}^{M} ||(G_p + X)\mathbf{u}_{p,i}||^2 =}$$

$$\boldsymbol{\sum_{i=l}^{M} \mathbf{u}_{p,i}^T (G_p + X)^T (G_p + X)\mathbf{u}_{p,i}}$$

$$\boldsymbol{\mathbf{x}^T B_p \mathbf{x} + 2\mathbf{w}_p^T \mathbf{x} + \sum_{i=1}^{M} \sigma_i^2(G_p) = f(\mathbf{x})}$$

and we want to minimize $\boldsymbol{f(\mathbf{x})}$.

Algorithm for $\mathbf{G_p}$

- In the original matrix we substitute $\mathbf{0}$ for missing value or write down the row average
- Compute the singular value of $\boldsymbol{G_p^T G_p}$.
- Construct $\boldsymbol{B_p}$, with entries $\boldsymbol{b_p(s,t)}$, in $\boldsymbol{(s,t)}$ places.

$$b_p(s,t) = \frac{1}{2}\sum_{i=l}^{m} v_{p,i}^T (F(i_s,j_s)^T F(i_t,j_t)$$

$$+F(i_t,j_t)^T F(i_s,j_s)) v_{p,i}.$$

- The matrix $\boldsymbol{B_p}$ gives the exact solution of $\boldsymbol{G_{p+1}}$.
- Solve for $\mathbf{x}$ in $\boldsymbol{B_p \cdot \mathbf{x}_{p+1} = -\mathbf{w}_p}$.
  $\boldsymbol{G_{p+1} := G_p + \mathbf{X}_{p+1}}$ until
  $\frac{\sum_{i=L}^{M} \boldsymbol{\sigma_i^2(G_{p+1})}}{\sum_{i=1}^{M} \boldsymbol{\sigma_i^2(G_{p+1})}}$ is very small ($\boldsymbol{L = \ell + 1}$).

6 IMPROVED FRAA (IFRAA)

Improved Fixed Rank Approximation Algorithm [7].

First use FRAA to find a completion G.

Then use a cluster algorithm,

(We used K-means repeating & refining cluster size),

to find a reasonable number of clusters of similar genes, each cluster is a relatively smaller matrix having an effective low rank.

For each cluster of genes apply FRAA separately to recover the missing entries in this cluster.

These results suggest that IFRAA has a potential for being an effective algorithm to recover blurred spots in digital images.

7 Clustering

Given a metric space $\mathbf{X}, d : \mathbf{X} \times \mathbf{X} \to \mathbb{R}_+$ and $\mathcal{X} := \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbf{X}$ are n distinct points, associate $M := (d(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^n \in \mathbb{R}_+^{n \times n}$.

Problem: Partition $\mathcal{X}$ to clusters $\mathcal{X} = \cup_{j=1}^m \mathcal{X}_j$ using M.

There are many different approaches to solve this problem.

K-Means Clustering

- **Divide genes randomly in K-clusters**
- **Compute centroid**
- **Compute genes distance to clusters**
- **Pick closest cluster**
- **Recompute new centroid**
- **Repeat until convergence**

IFRAA(FRAA +K-Means Clustering)

- **Impute G via FRAA**
- **Find clusters $\mathbf{c}_i$ using K-means.**
- **Compare $\mathbf{c}_i$ with G, remove data points corresponding to missing data.**
- **Impute applying FRAA to $\mathbf{c}_i$.**

8 SIMULATIONS 1

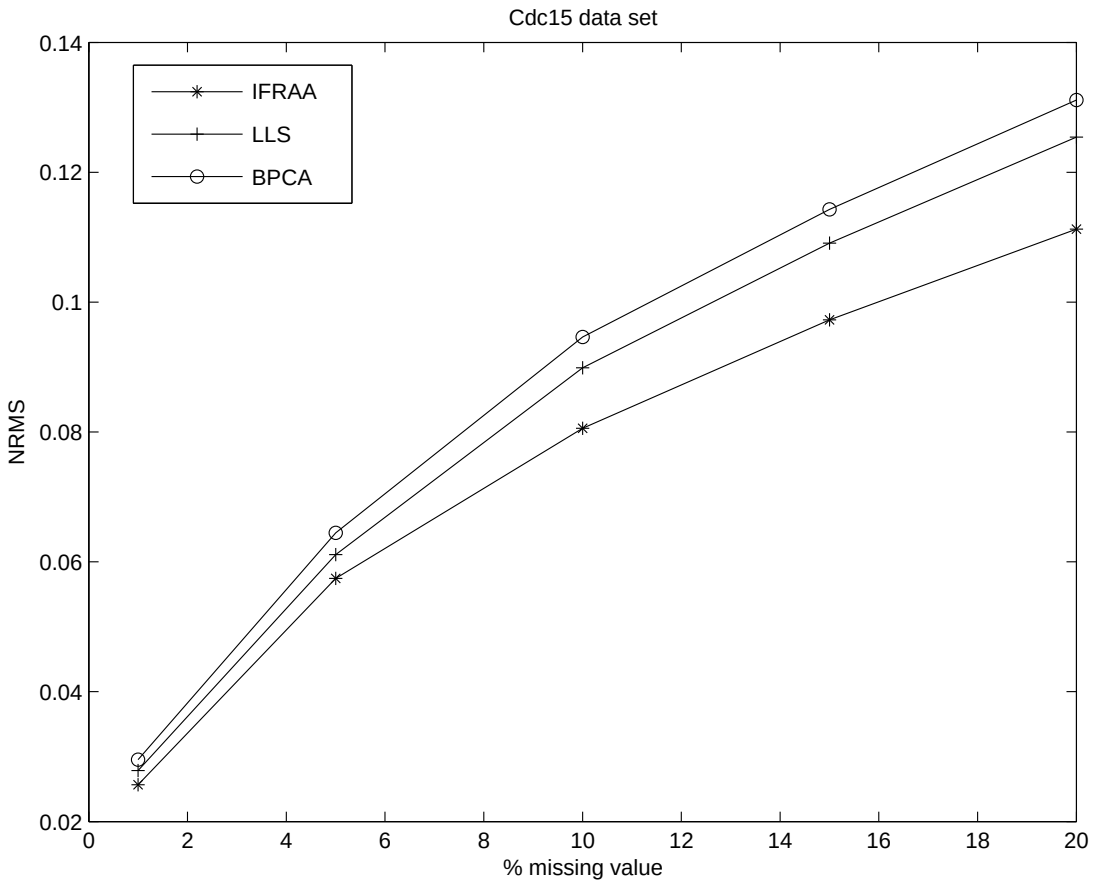

Figure 1: Comparison of NRMSE against percent of missing entries for three methods: IFRAA, BPCA and LLS. Cdc15 data set in [17] with 24 samples.

9 SIMULATIONS 2

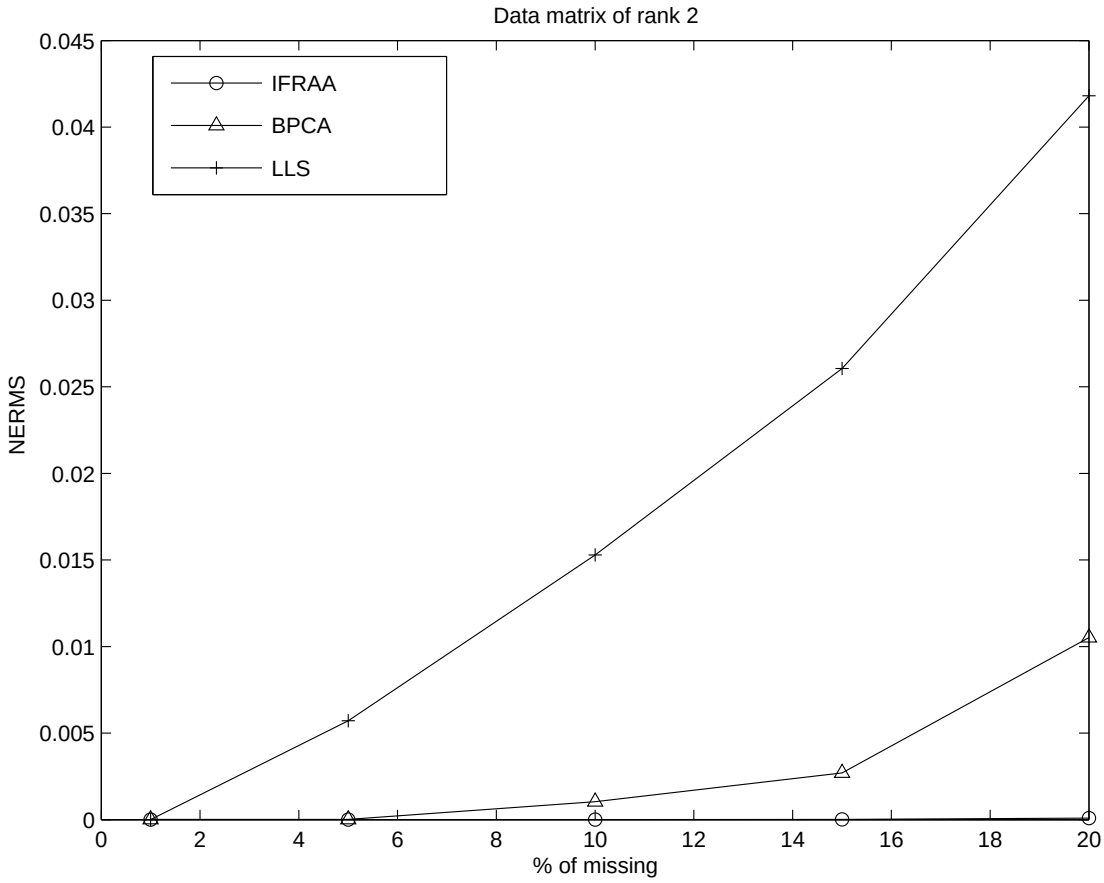

Figure 2: Comparison of NRMSE against percent of missing entries for three methods: IFRAA, BPCA and LLS. Data set was a $\mathbf{2000} \times \mathbf{20}$ randomly generated matrix of rank 2.

Bayesian principal component analysis-BPCA [15]: A global method consisting of three components. First, principal component regression, which is basically a low rank approximation of the data set is performed. Second, Bayesian estimation, which assumes that the residual error and the projection of each gene on principal components behave as normal independent random variables with unknown parameters, is carried out. Third, Bayesian estimation follows by iterations based on the expectation-maximization (EM) of the unknown Bayesian parameters.

Local least squares imputation method LLS [14]: A local methods, which use similarity structure of the data to impute the missing values. LLS has two versions to find similar genes whose expressions are not corrupted: the L_2-norm and the Pearson's correlation coefficients. After a group of similar genes C are identified, the missing values of the gene are obtained using least squares applied to the group C. The recovery of missing data is done independently, i.e. the estimation of each missing entry does not influence the estimation of the other missing entries.

10 TABLE

The performance of the BCPA, IFRAA and LLS algorithms depends on the unknown distribution of missing position of the entries.

Table 1: Comparison of NRMSE for three methods: IFRAA, LLS and BPCA for actual missing values distribution for three gene expression data sets with different percentage of missing values.

Data sets	IFRAA	LLS	BPCA
Cdc15 data set %0.81 missing	0.0175	0.0200	0.0216
Evolution data set %9.16	0.0703	0.0969	0.1247
Calcineurin data set %3.68	0.0421	0.0445	0.0453

11 Generalized SVD

Let $\boldsymbol{A} \in \mathbb{C}^{m\times n}, \boldsymbol{B} \in \mathbb{C}^{l\times n}$. Then Van Loan $\mathbf{70}\boldsymbol{s}$:
$\boldsymbol{A} = \boldsymbol{F}\boldsymbol{\Gamma}\boldsymbol{R},\ \boldsymbol{B} = \boldsymbol{G}\boldsymbol{\Delta}\boldsymbol{R},$
$\boldsymbol{F} \in \mathbf{U}(m),\ \boldsymbol{G} \in \mathbf{U}(l),\ \boldsymbol{R} \in \mathbf{GL}(n,\mathbb{C}),$
$\boldsymbol{\Gamma} \in \mathbb{R}_+^{m\times n}, \boldsymbol{\Delta} \in \mathbb{R}_+^{l\times n}$ diagonal matrices.

Numerical computations of GSVD are very unstable.

Thm ([5]). Let $\boldsymbol{P} := \boldsymbol{A}^*\boldsymbol{A} + \boldsymbol{B}^*\boldsymbol{B}$ and $r := \mathbf{rank}\,\boldsymbol{P}$.
Then $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Phi}\boldsymbol{V}^*, \boldsymbol{U} \in \mathbf{U}(m,r), \boldsymbol{V} \in \mathbb{C}^{n\times r},$
$\boldsymbol{B} = \boldsymbol{W}\boldsymbol{\Psi}\boldsymbol{V}^*, \boldsymbol{W} \in \mathbf{U}(l,r),$
$\boldsymbol{\Phi} = \mathbf{diag}(\phi_1,\ldots,\phi_r),$
$\boldsymbol{\Psi} = \mathbf{diag}(\psi_1,\ldots,\psi_r) \in \mathbb{R}_+^{r\times r}$ and
$\boldsymbol{\Phi}^2 + \boldsymbol{\Psi}^2 = \boldsymbol{I}_r.$

Hence $\boldsymbol{P} = \boldsymbol{V}\boldsymbol{V}^*$ and the columns of $\boldsymbol{V}$ form an orthonormal basis of the subspace $\mathbf{X}$, spanned by the columns of $\boldsymbol{A}^*, \boldsymbol{B}^*$ with respect to the inner product $\langle \mathbf{x}, \mathbf{y}\rangle := \mathbf{y}^*\boldsymbol{P}\mathbf{x}$ on $\mathbf{V}$.

Reason: $\boldsymbol{T}_{\boldsymbol{A}}^*\boldsymbol{T}_{\boldsymbol{A}} + \boldsymbol{T}_{\boldsymbol{B}}^*\boldsymbol{T}_{\boldsymbol{B}} = \boldsymbol{I}|_{\mathbf{X}} \Rightarrow$
$(\boldsymbol{T}_{\boldsymbol{A}}^*\boldsymbol{T}_{\boldsymbol{A}})(\boldsymbol{T}_{\boldsymbol{B}}^*\boldsymbol{T}_{\boldsymbol{B}}) = (\boldsymbol{T}_{\boldsymbol{B}}^*\boldsymbol{T}_{\boldsymbol{B}})(\boldsymbol{T}_{\boldsymbol{A}}^*\boldsymbol{T}_{\boldsymbol{A}})$

References

[1] O. Alter, P.O. Brown and D. Botstein, Singular value decomposition for genome-wide expression data processing and modelling, *Proc. Nat. Acad. Sci. USA* 97 (2000), 10101-10106.

[2] 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.

[3] 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, to appear.

[4] A. Deshpande, L. Rademacher, S. Vemapala and G. Wang, Matrix Approximation and Projective Clustering via Volume Sampling, *SODA*, 2006.

[5] S. Friedland, A New Approach to Generalized Singular Value Decomposition, SIMAX 27 (2005), 434-444.

[6] S. Friedland, M. Kaveh, A. Niknejad and H. Zare, Fast Monte-Carlo Low Rank Approximations for Matrices, Proc. IEEE SoSE, 2006.

[7] S. Friedland, M. Kaveh, A. Niknejad and H. Zare, An Algorithm for Missing Value Estimation for DNA Microarray Data, Proc. ICASSP, 2006

[8] S. Friedland, A. Niknejad and L. Chihara, A simultaneous reconstruction of missing data in DNA microarrays, to appear in *Linear Algebra and Its Applications*.

[9] 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.

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

[11] G.H. Golub and C.F. Van Loan, *Matrix Computation*, John Hopkins Univ. Press, 3rd Ed., 1996.

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

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

[14] 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.

[15] 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.

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

[17] 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.

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