A Simultaneous Reconstruction of Missing Data in DNA Microarrays

Shmuel Friedland [*†], Amir Niknejad [*], and Laura Chihara[‡]
Institute for Mathematics and its Applications
400 Lind Hall, 207 Church St., S.E.
Minneapolis, MN 55455-0436

December 1, 2003

Abstract

We suggest here a new method of the estimation of missing entries in a gene expression matrix, which is done simultaneously— i.e., the estimation of one missing entry influences the estimation of other entries. Our method is closely related to the methods and techniques used for solving inverse eigenvalue problems.

2000 Mathematical Subject Classification: 15A18, 92D10
Keywords: Gene expression matrix, singular value decomposition (svd), missing values, imputation

1 Introduction

In the last decade, molecular biologists have been using DNA microarrays as a tool for analyzing information in gene expression data. During the laboratory process, some spots on the array may be missing due to various factors (for example, machine error.) Because it is often very costly or time consuming to repeat the experiment, molecular biologists, statisticians, and computer scientists have made attempts to recover the missing gene expressions by some ad-hoc and systematic methods.

[*] *Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago, Chicago, Illinois 60607-7045*
[†] Corresponding author. Tel.: +1-312-996-3041; fax: +1-312-996-1491; email: friedlan@uic.edu
[‡] *Department of Mathematics and Computer Science, Carleton College, One N. College St., Northfield, MN 55057*

More recently, microarray gene expression data have been formulated as a gene expression matrix E with N rows, which correspond to genes, and M columns, which correspond to experiments. Typically N is much larger than M. In this setting, the analysis of missing gene expressions on the array would translate to recovering missing entries in the gene expression matrix.

The most common methods for recovery are [8]:
(a) Clustering analysis methods such as K-nearest neighbor clustering , hierarchical clustering, or
(b) SVD - Singular Value Decomposition (also known as Principal Component Analysis).

In these methods, 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. The iterative method using SVD suggested in [8] takes into account implicitly the influence of the estimation of one entry on the other ones. See also [2].

We suggest a new method in which the estimation of missing entries is done simultaneously, i.e., the estimation of one missing entry influences the estimation of the other missing entries. If the gene expression matrix E has missing data, we want to complete its entries to obtain a matrix $\hat{E}$, such that the rank of $\hat{E}$ is equal to (or does not exceed) d, where d is taken to be the number of significant singular values of E. The estimation of the entries of E to a matrix with a prescribed rank is a variation of the *problem of communality* (*see* [4, p. 637].) We give an optimization algorithm for finding $\hat{E}$ using the techniques for inverse eigenvalue problems discussed in [3].

We implemented our fixed rank approximation algorithm as a Matlab procedure and ran simulations on the microarray data *Saccharomyces cerevisiae* [7]. We describe the results in Section 7.

2 The Singular Value Decomposition

In this section, we recall some basic facts about *Singular Value Decomposition SVD*. Let E be an $N \times M$ real-valued nonzero matrix. In this paper we assume that $N \geq M$. The SVD of E is a decomposition of E into the product $U\Sigma V^{\mathrm{T}}$ with certain properties. There are a few variations of this definition, and we give the following one which is most suitable for the applications in our context. We assume that U is $N \times M$, Σ is $M \times M$, and V is $M \times M$.

$$E = U\Sigma V^{\mathrm{T}},\ U^{\mathrm{T}}U = V^{\mathrm{T}}V = I_M,\ \Sigma = \mathrm{diag}(\sigma_1, ..., \sigma_M),\ \sigma_1 \geq ... \geq \sigma_M \geq 0. \quad (2.1)$$

The rank r of E is the number of positive singular values; the dimension of the row space, and the dimension of the column space of E is also r.

Remark Singular value decomposition is related to principal component analysis (PCA) in statistics. If we center each column in matrix E, then $E^TE = V\Sigma^2V^T$ is proportional to the covariance matrix of the columns of E, the columns of V are the principal components, and the $\{\sigma_i^2\}$ are proportional to the variances of the principal components.

Let U_r, Σ_r, V_r be matrices obtained from U, Σ, V, respectively, as follows: U_r is an $N \times r$ matrix obtained by deleting the last $N - r$ columns of U, V_r is the $M \times r$ matrix obtained by deleting the last $M - r$ columns of V, and Σ_r is obtained by deleting the last $M - r$ columns and rows of Σ . Then

$$E = U_r\Sigma_r V_r^{\mathrm{T}},\ U_r^{\mathrm{T}}U_r = V_r^{\mathrm{T}}V_r = I_r,\ \Sigma_r = \mathrm{diag}(\sigma_1, ..., \sigma_r),\ \sigma_1 \geq ... \geq \sigma_r > 0. \quad (2.2)$$

In this setting U_r, Σ_r, V_r are all rank r matrices: the last $N-r$ columns of U and the last $M-r$ rows of V^{T} are arbitrary, up to the condition that the last $N-r$ columns U and last $M-r$ rows of V^{T} are orthonormal bases of the orthogonal complement of the column space and the row space of E respectively. Hence (2.2) is sometimes called a *reduced* SVD of E.

We now give another form of (2.2) which has a significant interpretation in microarray data. Let $\mathbf{u}_1, \cdots, \mathbf{u}_M$ denote the columns of U and $\mathbf{v}_1, \ldots, \mathbf{v}_M$ denote the columns of V. Then (2.1) and (2.2) can be written as

$$E = \sum_{i=1}^{M} \sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}}. \quad (2.3)$$

If $\sigma_1 > ... > \sigma_r$ then $\mathbf{u}_i$ and $\mathbf{v}_i$ are determined up to the sign ± 1 for $i = 1, ..., r$. Namely $\mathbf{u}_i$ and $\mathbf{v}_i$ are length 1 eigenvectors of EE^{T} and $E^{\mathrm{T}}E$, respectively, corresponding to the common eigenvalue σ_i^2. (Note the choice of a sign in $\mathbf{v}_i$ forces a unique choice of the sign in $\mathbf{u}_i$.) Computationally, one first finds the positive eigenvalues $\sigma_1^2, ..., \sigma_r^2$ and the corresponding orthonormal eigenvectors $\mathbf{v}_1, ..., \mathbf{v}_r$ of the smaller matrix $E^{\mathrm{T}}E$. Then

$$\mathbf{u}_i := \frac{1}{\sigma_i} E\mathbf{v}_i \iff \sigma_i \mathbf{u}_i = E\mathbf{v}_i, \quad i = 1, ..., r. \quad (2.4)$$

To compute the decomposition (2.3), it is enough to know $\mathbf{v}_i$ and $\sigma_i\mathbf{u}_i$. If σ_i repeats $k > 1$ times in the sequence $\sigma_1 \geq ... \geq \sigma_r > 0$, then the choice of the corresponding k eigenvectors $\mathbf{v}_j$ is not unique: any choice of the orthonormal basis in the eigenspace of $E^{\mathrm{T}}E$ corresponding to the eigenvalue σ_i^2 is a legitimate choice.

Denote by $||E||_{\mathcal{F}}$ the Frobenius (ℓ_2) norm of E. It is the Euclidean norm of E viewed as a vector with NM coordinates. Each term $\mathbf{u}_i\mathbf{v}_i^{\mathrm{T}}$ in (2.3) is a rank one

matrix with $||\mathbf{u}_i\mathbf{v}_i^{\mathrm{T}}||_{\mathcal{F}} = 1$. 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:

$$\min_{F \in \mathcal{R}(N,M,k)} ||E - F||_{\mathcal{F}} = ||E - \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}}||_{\mathcal{F}} = \sqrt{\sum_{i=k+1}^{r} \sigma_i^2}. \tag{2.5}$$

If $\sigma_k > \sigma_{k+1}$ then $\sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}}$ is the unique solution to the above minima problem. For the purposes of this paper, it will be convenient to assume that $\sigma_i = 0$ for any $i > M$.

In what follows we will use yet another equivalent definition of the singular values of E. Let $\mathcal{M}_{NM}$ denote the space of all real $N \times M$ matrices and S_J denote the space of all real $J \times J$ symmetric matrices. For $A \in S_J$, we let

$$\lambda_1(A) = \lambda_1 \geq ... \geq \lambda_J(A) = \lambda_J,\ A\mathbf{w}_i = \lambda_i \mathbf{w}_i,\ \mathbf{w}_i^{\mathrm{T}}\mathbf{w}_j = \delta_{ij},\ i, j = 1, ..., J, \tag{2.6}$$

be the eigenvalues and corresponding eigenvectors of A, where the eigenvalues are counted with their multiplicities, and the eigenvectors form an orthonormal basis in $\mathbb{R}^J$.

Consider the following $(N + M) \times (N + M)$ real symmetric matrix:

$$S(E) := \begin{pmatrix} 0 & E \\ E^{\mathrm{T}} & 0 \end{pmatrix}. \tag{2.7}$$

It is known [5, §7.3.7]

$$\begin{aligned} &\sigma_i(E) := \sigma_i = \lambda_i(S(E)) = -\lambda_{N+M+1-i}(S(E)), \text{ for } i = 1, ..., M, \\ &\lambda_i(S(E)) = 0 \text{ for } i = M + 1, ..., N. \end{aligned} \tag{2.8}$$

The Cauchy interlacing property for $S(E)$ implies [5, §7.3.9]

Let $[N] := \{1, 2, \ldots, N\}$, and let $\mathcal{N} \subset [N]$, $\mathcal{M} \subset [M]$ denote sets of cardinalities $N', M' \geq 0$ respectively.

Proposition 2.1 *Let $E \in \mathcal{M}_{NM}$ and denote by $E' \in \mathcal{M}_{(N-N')(M-M')}$ the matrix obtained from E by deleting all rows $n \in \mathcal{N}$ and all columns $m \in \mathcal{M}$. Then*

$$\begin{aligned} &\sigma_i(E) \geq \sigma_i(E') \text{ for } i = 1, ..., M, \\ &\sigma_i(E') \geq \sigma_{i+N'+M'}(E) \text{ for } i = 1, ..., M - (M' + N'). \end{aligned} \tag{2.9}$$

3 The Gene Expression Matrix

In this section we will view $E \in \mathcal{M}_{NM}$, with $N \geq M$ as the gene expression matrix:

$$E = \begin{pmatrix} g_{11} & g_{12} & \cdots & g_{1M} \\ g_{21} & g_{22} & \cdots & g_{2M} \\ \vdots & \vdots & \vdots & \vdots \\ g_{n1} & g_{n2} & \cdots & g_{nM} \\ \vdots & \vdots & \vdots & \vdots \\ g_{N1} & g_{N2} & \cdots & g_{NM} \end{pmatrix} = \begin{pmatrix} \mathbf{g}_1^{\mathrm{T}} \\ \mathbf{g}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{g}_n^{\mathrm{T}} \\ \vdots \\ \mathbf{g}_N^{\mathrm{T}} \end{pmatrix} = [\, \mathbf{c}_1 \quad \mathbf{c}_2 \quad \ldots \mathbf{c}_M \,], \tag{3.1}$$

$$\mathbf{g}_n^{\mathrm{T}} := (g_{n1}, g_{n2}, ..., g_{nM}),\ n = 1, ..., N, \quad \mathbf{c}_i = \begin{pmatrix} g_{1i} \\ g_{2i} \\ \vdots \\ g_{ni} \\ \vdots \\ g_{Ni} \end{pmatrix},\ i = 1, ..., M.$$

The row vector $\mathbf{g}_n^{\mathrm{T}}$ corresponds to the (relative) expression levels of the n^{th} gene in M experiments. The column vector $\mathbf{c}_i$ corresponds to the (relative) expression levels of the N genes in the i^{th} experiment.

Consider the SVD of the gene expression matrix $E = U\Sigma V^T$. In the terminology of [1], the columns of U are eigenarrays, the columns of V are eigengenes, and the singular values of E are eigenexpression levels.

In many microarray data sets, researchers have found that only a few eigengenes are needed to capture the overall gene expression pattern. The number of these *significant* eigengenes is determined heuristically. For example, set

$$p_i := \frac{\sigma_i^2}{\sum_{j=1}^{M} \sigma_j^2},\ i = 1, ..., M, \quad \mathbf{p} := (p_1, ..., p_M)^{\mathrm{T}}, \tag{3.2}$$

so that p_i represents the fraction of the expression level contributed by the i^{th} eigengene. Then we choose the L eigengenes that contribute about $70\% - 90\%$ of the total expression level. Another method is to use scree plots for the σ_i^2. (In principal component analysis, the p_i are proportional to the variances of the principal components, so we choose the principal components of maximum variability [6].)

If E has L significant eigenvalues, we view σ_i to be effectively equal to zero for $i > L$. We define the matrix

$$E_L := \sum_{i=1}^{L} \sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}} \tag{3.3}$$

as the *filtered* part of E and consider $E - E_L$ the *noise* part of E.

Let

$$1 \geq h(\mathbf{p}) := -\frac{1}{\log M} \sum_{i=1}^{M} p_i \log p_i \geq 0. \tag{3.4}$$

Then $h(\mathbf{p})$ is the rescaled entropy of the probability vector $\mathbf{p}$. $h(\mathbf{p}) = 1$ only when $p_i = \frac{1}{M},\ i = 1, ..., M$; in other words, all the eigengenes are equally expressed. On the other hand, $h(\mathbf{p}) = 0$ if and only if $p_i(1 - p_i) = 0, i = 1, ..., M$ and this corresponds to $r = 1$: in other words, the gene expression is captured by a single eigengene (and eigenarray).

The following example points out a potential weakness of SVD theory in trying to detect groups of genes with similar properties.

3.1 SVD and gene clusters

Suppose the set of genes $\mathbf{g}_n^{\mathrm{T}}$, $n \in [N]$ can be grouped into $K+1$ disjoint subsets $[N] = \cup_{j=1}^{K+1} G_j$ with $G_1, ..., G_K$ nonempty and $M \geq K$ (usually $M > K$). In particular, consider the genes in each group G_j ($j = 1, ..., K$) to have similar characteristics (in other words, G_j is a cluster). Genes that have no similar characteristics are placed in G_{K+1}. Also,

$$\begin{aligned} &g_{pj} = a_{kj} \text{ for each } p \in G_k \text{ and } k = 1, ..., K,\ j = 1, ..., M, \\ &g_{pj} = 0 \text{ for each } p \in G_{K+1} \text{ and } j = 1, ..., M, \end{aligned} \tag{3.5}$$

Let $A = (a_{kj})_{k,j=1}^{K,M} \in \mathrm{M}_{KM}$ be the corresponding $K \times M$ matrix with the rows $\mathbf{r}_1^{\mathrm{T}}, ..., \mathbf{r}_L^{\mathrm{T}}$:

$$A = \begin{pmatrix} \mathbf{r}_1^{\mathrm{T}} \\ \mathbf{r}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{r}_K^{\mathrm{T}} \end{pmatrix}.$$

Clearly the row space of E is the row space of A. So $K \geq \operatorname{rank} E = \operatorname{rank} A$. Hence if $\operatorname{rank} A = K$ then

$$\sigma_1(E) \geq ... \geq \sigma_K(E) > \sigma_{K+1}(E) = ... = \sigma_M(E) = 0.$$

However, there is no simple formula relating the singular values of E and A. It may happen that the rows of A are linearly dependent which indicates that several groups out of $G_1, ..., G_K$ are somehow related, and the number of the significant singular values of E is less than K.

4 Missing Data in the Gene Expression Matrix

We now consider the problem of missing data in the gene expression matrix E. (Our analysis can be applied to any matrix E.) Let $\mathcal{N} \subset [N]$ denote the set of rows of E that contain at least one missing entry. Thus for each $n \in \mathcal{N}^c := [N]\backslash\mathcal{N}$, the gene $\mathbf{g}_n^{\mathrm{T}}$ has all of its entries. Let N' denote the size of $\mathcal{N}^c$ so that the size of $\mathcal{N}$ is $N - N'$. We want to complete the missing entries of each $\mathbf{g}_n^{\mathrm{T}}$, $n \in \mathcal{N}$, under some assumptions.

We first describe the reconstruction of the missing data in E using SVD as given in [1].

4.1 Imputation using SVD

Let E' be the $N' \times M$ matrix containing the rows $\mathbf{g}_m^{\mathrm{T}}$, $m \in \mathcal{N}^c$ of E which do not have any missing entries, and L' be the number of significant singular values of E'. Let $\mathbf{X} \subset \mathbb{R}^M$ be the invariant subspace of the symmetric matrix $(E')^{\mathrm{T}}E'$ corresponding to the eigenvalues $\sigma_1(E')^2, ..., \sigma_{L'}(E')^2$. Let $\mathbf{x}_1, ..., \mathbf{x}_{L'}$ be the orthonormal eigenvectors of $(E')^{\mathrm{T}}E$ corresponding to the eigenvalues $\sigma_1(E')^2, ..., \sigma_{L'}(E')^2$. Then $\mathbf{x}_1, ..\mathbf{x}_{L'}$ is a basis of $\mathbf{X}$.

Let $\mathcal{M} \subset [M]$ be a subset of cardinality M'. Consider the projection $\pi_{\mathcal{M}} : \mathbb{R}^M \to \mathbb{R}^{M-M'}$ by deleting the coordinate $i \in \mathcal{M}$ for any vector $\mathbf{x} = (x_1, ..., x_M)^{\mathrm{T}} \in \mathbb{R}^M$. Then $\pi_{\mathcal{M}}(\mathbf{X})$ is spanned by $\pi_{\mathcal{M}}(\mathbf{x}_1), ..., \pi_{\mathcal{M}}(\mathbf{x}_{L'})$.

Fix $n \in \mathcal{N}$ and let $\mathcal{M} \subset [M]$ be the set of experiments (columns) where the gene $\mathbf{g}_n^{\mathrm{T}}$ has missing entries. Let $\mathbf{y} \in \pi_{\mathcal{M}}(\mathbf{X})$ be the least square approximation to $\pi_{\mathcal{M}}(\mathbf{g}_n)$. Then any $\overline{\mathbf{g}}_n \in \pi_{\mathcal{M}}^{-1}(\mathbf{y})$ is a completion of $\mathbf{g}_n$. If $\pi_{\mathcal{M}}|\mathbf{X}$ is 1-1 then $\overline{\mathbf{g}}_n$ is unique. Otherwise one can choose $\overline{\mathbf{g}}_n \in \pi_{\mathcal{M}}^{-1}(\mathbf{y})$ with the least norm. Note that to find $\mathbf{y} \in \pi_{\mathcal{M}}(\mathbf{X})$ one needs to solve the least square problem for a subspace $\pi_{\mathcal{M}}(\mathbf{X})$. In principle, for each $n \in \mathcal{N}$ one solves a different least square problem. The crucial assumption of this method that

$$L = L'. \tag{4.1}$$

The significant singular values of E' and of the reconstructed E are joint functions of all the rows (genes). By trying to reconstruct the missing data in each gene $\mathbf{g}_n^{\mathrm{T}}$, for $n \in \mathcal{N}$, separately, we ignore any correlation between $\mathbf{g}_n^{\mathrm{T}}$ and the genes $\mathbf{g}_k^{\mathrm{T}}$, $k \in \mathcal{N}$; consequently, this will have an impact on the singular values of E. In the following section we suggest a different approach which treats the estimation problem of all the missing data simultaneously.

4.2 Reconsideration of 3.1

Let us reconsider Example 3.1. Assume that rank $A = K$. Then we can reconstruct exactly each missing entry of $\mathbf{g}_n^{\mathrm{T}},\ n \in \mathcal{N}$ if and only if $G_i \backslash \mathcal{N} \neq \emptyset$ for $i = 1, ..., M$. In this example this condition is equivalent to the assumption that E' has the same rank as E.

4.3 Iterative method using SVD

In the recent papers [8] and [2], the following iterative method using SVD to impute missing values in a gene expression matrix is suggested. First, replace the missing values with 0 or with values computed from another method. Call the estimated matrix E_p, where $p = 0$. Find the L_p significant singular values of E_p, and let E_{L_p} be the filtered part of E_p (3.3). Replace the missing values in E by the corresponding values in E_{L_p} to obtain the matrix E_{p+1}. Continue this process until E_p converges to a fixed matrix (within a given precision). This algorithm takes into account implicitly the influence of the estimation of one entry on the other ones. But it is not clear if the algorithm converges, nor what are the features of any fixed point(s) of this algorithm.

5 The Optimization Problem

We now show that the estimation problem discussed in the previous section can be cast as the following optimization problem:

Problem 5.1 *Let $\mathcal{S} \subset [N] \times [M]$ and denote by $E(\mathcal{S})$ a given set of real numbers e_{ij} for $(i, j) \in \mathcal{S}$. Let $\mathrm{M}(E(\mathcal{S})) \subset \mathcal{M}_{NM}$ be the affine subset of all matrices $A = (a_{ij}) \in \mathcal{M}_{NM}$ such that $a_{ij} = e_{ij}$ for all $(i, j) \in \mathcal{S}$. Let ℓ be a positive integer not exceeding M. Find $\hat{E} \in \mathrm{M}(E(\mathcal{S}))$ with the minimal σ_ℓ.*

Let $E = (g_{ij})$ denote the gene expression matrix with missing values. We choose the $\mathcal{S}$ in Problem 5.1 to be the set of coordinates coordinates (i, j) for which the entry g_{ij} is not missing. Recall that $\mathcal{N} \subset [N]$ denotes the set of rows of E that contain at least one missing entry. Hence the set $\mathcal{S}$ contains all the rows $i \in \mathcal{N}^c$. Clearly, the complement of $\mathcal{S}$ is the set of coordinates $\mathcal{S}^c = \{(i, j) \,|\, g_{ij} \text{ is missing}\} \subset \mathcal{N} \times [M]$. Let N_1 denote the total number of missing entries in E; thus $N_1 \geq N'$.

Let E' be the matrix as in 4.1 with L' significant singular values. Note that (2.9) yields $\sigma_i(E) \geq \sigma_i(E')$ for $i = 1, ..., M$. Thus if we want to complete E such that the resulting matrix still has exactly L' significant singular values, we should consider Problem 5.1 with $\ell = L' + 1$.

A more general possibility is to assume that the number of significant singular values of a possible estimation of E is $L = L' + k$ where k is a small integer, e.g. $k = 1$ or 2. That is, in the group of genes $\mathbf{g}_n^{\mathrm{T}},\ n \in \mathbb{N}$, there are k significant genes which are not found in the group of genes in $\mathcal{N}^c$. Then one considers Problem 5.1 with $\ell = L' + k + 1$.

We now consider a modification of Problem 5.1 which has a nice numerical algorithm.

Problem 5.2 *Let $\mathcal{S} \subset [N] \times [M]$ and denote by $E(\mathcal{S})$ a given set of real numbers e_{ij} for $(i,j) \in \mathcal{S}$. Let $\mathrm{M}(E(\mathcal{S})) \subset \mathrm{M}_{NM}$ the affine subset of all matrices $A = (a_{ij}) \in \mathrm{M}_{NM}$ such that $a_{ij} = e_{ij}$ for all $(i,j) \in \mathcal{S}$. Let ℓ be a positive integer not exceeding M. Find $\hat{E} \in \mathrm{M}(E(\mathcal{S}))$ such that $\sum_{i=\ell}^{M} \sigma_i^2$ is minimal.*

Clearly, we can find $E \in \mathrm{M}(E(\mathcal{S}))$ with a "small" $\sigma_\ell^2(E)$ if and only if we can find $E \in \mathrm{M}(E(\mathcal{S}))$ with a "small" $\sum_{i=\ell}^{M} \sigma_i^2(E)$.

6 Fixed Rank Approximation Algorithm

We now describe one of the standard algorithms to solve Problem 5.2.

Algorithm 6.1 Fixed Rank Approximation Algorithm (FRAA) *Let $E_p \in \mathrm{M}(E(\mathcal{S}))$ be the p^{th} approximation to a solution of Problem 5.2. Let $A_p := E_p^{\mathrm{T}} E_p$ and find an orthonormal set of eigenvectors for A_p, $\mathbf{v}_{p,1}, ..., \mathbf{v}_{p,M}$ as in (2.6). Then E_{p+1} is a solution to the following minimum of a convex nonnegative quadratic function*

$$\min_{E \in \mathrm{M}(E(\mathcal{S}))} \sum_{i=\ell}^{M} (E\mathbf{v}_{p,i})^{\mathrm{T}} (E\mathbf{v}_{p,i}). \tag{6.1}$$

We now explain the algorithm and show that in each step, we decrease the value of the function we minimize:

$$\sum_{i=\ell}^{M} \sigma_i^2(E_p) \geq \sum_{i=\ell}^{M} \sigma_i^2(E_{p+1}). \tag{6.2}$$

For any integer $k \in [M]$, let Ω_k denote the set of all k orthonormal vectors $\{\mathbf{y}_1, ..., \mathbf{y}_k\}$ in $\mathbb{R}^M$. Let $A \in \mathcal{S}_M$ and assume (2.6) with $J = M$. Then the minimal principle (the Ky-Fan characterization for $-A$) is:

$$\sum_{i=\ell}^{M} \lambda_i(A) = \sum_{i=\ell}^{M} \mathbf{w}_i^{\mathrm{T}} A \mathbf{w}_i = \min_{\{\mathbf{y}_\ell, ..., \mathbf{y}_M\} \in \Omega_{M-\ell+1}} \sum_{i=\ell}^{M} \mathbf{y}_i^{\mathrm{T}} A \mathbf{y}_i. \tag{6.3}$$

See for example [3].

Let $E = E_p + X \in \mathrm{M}(E(\mathcal{S}))$. Then $X = (x_{i,j})_{i,j=1}^{N,M}$ where $x_{ij} = 0$ if $(i,j) \in \mathcal{S}$ and x_{ij} is a free variable if $(i,j) \notin \mathcal{S}$.

Let $\mathbf{x} = (x_{i_1,j_1}, x_{i_2,j_2}, \ldots, x_{i_{N_1},j_{N_1}})^T$ denote the $N_1 \times 1$ vector whose entries are indexed by $\mathcal{S}^c$, the coordinates of the missing values in E. Then there exists a unique $N_1 \times N_1$ real valued symmetric nonnegative definite matrix $N_1 \times N_1$ matrix B_p which satisfies the equality

$$\mathbf{x}^{\mathrm{T}} B_p \mathbf{x}^{\mathrm{T}} = \sum_{i=\ell}^{M} \mathbf{u}_{p,i} X^{\mathrm{T}} X \mathbf{u}_{p,i}. \tag{6.4}$$

Let $F(i,j)$ be the $N \times M$ matrix with 1 in the (i,j) entry and 0 elsewhere. Then the (s,t) entry of B_p is given by

$$b_p(s,t) = \frac{1}{2} \sum_{i=\ell}^{M} \mathbf{v}_{p,i}^{\mathrm{T}} (F(i_s,j_s)^{\mathrm{T}} F(i_t,j_t) + F(i_t,j_t)^{\mathrm{T}} F(i_s,j_s)) \mathbf{v}_{p,i}, \tag{6.5}$$
$$s,t = 1, \ldots N_1$$

The proof of 6.5 is given in the Appendix. The crucial observation is that B_p can be decomposed into the direct sum of N_1 symmetric nonnegative definite matrices indexed by $\mathcal{N}$.

Hence the function minimized in (6.1) is given by

$$\sum_{i=\ell}^{M} \mathbf{v}_{p,i}^{\mathrm{T}} E^{\mathrm{T}} E \mathbf{v}_{p,i} = \sum_{i=\ell}^{M} \mathbf{v}_{p,i}^{\mathrm{T}} (A_p + E_p^{\mathrm{T}} X + X^{\mathrm{T}} E_p + X^{\mathrm{T}} X) \mathbf{v}_{p,i} =$$
$$\mathbf{x}^{\mathrm{T}} B_p \mathbf{x} + 2\mathbf{w}_p^{\mathrm{T}} \mathbf{x} + \sum_{i=\ell}^{M} \lambda_i(A_p) =$$
$$\sum_{i \in \mathcal{N}} (\mathbf{x}_i^{\mathrm{T}} B_{p,i} \mathbf{x}_i + 2\mathbf{w}_{p,i}^{\mathrm{T}} \mathbf{x}_i) + \sum_{i=\ell}^{M} \lambda_i(A_p), \tag{6.6}$$
$$\text{where } \mathbf{w}_p := (w_{p,1}, \ldots, w_{p,N_1})^{\mathrm{T}}, \text{ for } i \in \mathcal{N},$$
$$w_{p,t} = \sum_{i=\ell}^{M} \mathbf{v}_{p,i}^{\mathrm{T}} E_p^{\mathrm{T}} F(i_t,j_t) \mathbf{v}_{p,i}, \ t = 1, ..., N_1.$$

Since the expression in (6.1), and hence in (6.6), is always nonnegative, it follows that $\mathbf{w}_p$ is in the column space of B_p. Hence the minimum of the function given in (6.6) is achieved at the critical point

$$B_p \mathbf{x}_{p+1} = -\mathbf{w}_p, \tag{6.7}$$

and this system of equations is always solvable. (If B_p is not invertible, we find the least-squares solution).

We now show (6.2). The vector $\mathbf{x}_{p+1}$ contains the entries for the matrix X_{p+1}. Then $E_{p+1} := E_p + X_{p+1}$. From the definition of $A_{p+1} := E_{p+1}^{\mathrm{T}} E_{p+1}$ and the minimality of $\mathbf{x}_{p+1}$ we obtain

$$\sum_{i=\ell}^{M} \sigma_i(E_p)^2 = \sum_{i=\ell}^{M} \mathbf{v}_{p,i}^{\mathrm{T}} (E_p + 0)^{\mathrm{T}} (E_p + 0) \mathbf{v}_{p,i} \geq$$
$$\sum_{i=\ell}^{M} \mathbf{v}_{p,i}^{\mathrm{T}} (E_p + X_{p+1})^{\mathrm{T}} (E_p + X_{p+1}) \mathbf{v}_{p,i} = \sum_{i=\ell}^{M} \mathbf{v}_{p,i}^{\mathrm{T}} A_{p+1} \mathbf{v}_{p,i} \geq$$
$$\sum_{i=\ell}^{M} \lambda_i(A_{p+1}) = \sum_{i=\ell}^{M} \sigma_i(E_{p+1})^2.$$

□

In Appendix B, we give an algorithm to solve 6.7 efficiently. We conclude this section by remarking that to solve Problem 5.1, one may use the methods of [4].

7 Simulation

We implemented the Fixed Rank Approximation Algorithm (FRAA) in Matlab and tested it on the microarray data *Saccharomyces cerevisiae* [7] as provided at http://genome-www.stanford.edu (the elutriation data set). The dimensions of the complete gene expression matrix is 5981×14. We randomly deleted a set of entries and ran FRAA on this "corrupted" matrix to obtain estimates for the deleted entries. The FRAA requires four inputs: the matrix with missing entries, an initial guess, a parameter L–the number of significant singular values, and the number of iterations. We set the initial guess to the missing data matrix with 0's replacing the missing values, the number of significant values to $L = 2$, and ran the algorithm through 5 iterations.

We compared our estimates to estimates obtained by three other methods: replacing missing values with 0's (zeros method), row means (row means method), or the values obtained by the KNNimpute program [8]. We used a normalized root mean square as the metric for comparison: if C represents the complete matrix and E_p represents an estimate to the corrupted matrix E, then the root mean square (RMS) of the difference $D = C - E_p$ is $\frac{||D||_{\mathcal{F}}}{\sqrt{N}}$, where N is the length (the larger of

the two dimensions) of D. We normalized the root mean square by dividing RMS by the average value of the entries in C.

In simulations where $1\% - 20\%$ of the entries were randomly deleted from the complete matrix C, the FRAA performed slightly better than the row means method, and significantly better than the zeros method. However, the KNNimpute algorithm (with parameters k=15, d=0) produced the most accurate estimates, with normalized RMS errors that were smaller than the normalized RMS errors from the other three methods. Figure 7.1 gives the results of one set of experiments: the normalized RMS errors is plotted against percent missing. Not surprisingly, normalized RMS's increase with increasing percentage of missing values.

In [8], the authors caution against using KNNimpute for matrices with fewer than 6 columns. We randomly selected four columns from the elutriation data set to form a truncated data set, then randomly deleted from $1\% - 20\%$ of the entries from this newly formed matrix. Figure 7.2 gives a comparison of the normalized RMS errors against percent missing for three of the estimation methods. In this case, FRAA performed slightly better than the KNNimpute algorithm. Figure 7.2(b) shows the distribution of raw errors (true value - estimated value) for one simulation where 10% (2400) of the entries were deleted and then estimated. The standard deviation of the errors was .206 for FRAA, .237 for KNNimpute, and .255 for the row means method.

However, in other simulations choosing four other different columns from C, the results were mixed: sometimes FRAA gave the smallest normalized RMS, other times KNNimpute gave the smallest normalized RMS. In all cases, these two methods were more accurate than the row means method for imputation.

8 Discussion

The Fixed Rank Approximation Algorithm uses singular value decomposition to obtain estimates of missing values in a gene expression matrix. It uses all the known information in the matrix to simultaneously estimate all missing entries. Preliminary tests indicate that FRAA is more accurate than replacing missing values with 0's or with row means. The KNNimpute algorithm was more accurate when estimating missing entries deleted from the full elutriation matrix, but FRAA might be a feasible alternative in cases when the number of columns is small.

FRAA is another option for estimating missing values in gene expression data. Future work should look at estimating missing data from other types of microarray data sets. The biology of the data should guide the researcher in determining the best method to use for imputing missing values in these data sets.

Appendix

A Proof of 6.5

Let $\mathcal{N} \subset [N]$. Let $\mathcal{S}(i)$ denote the set of coordinates in row i with known values in E so that $\mathcal{S}(i)^{\mathbf{c}}$ denotes the set of coordinates of the missing values in row i.

$$\mathcal{S}^c = \cup_{i\in\mathcal{N}}\mathcal{S}(i)^c,\ \mathcal{S}(i)^c = \{(i,j(i,1)),...,(i,j(i,k(i)))\}, \tag{A.1}$$
$$M \geq j(i,k(i)) > ... > j(i,1) \geq 1 \quad \text{for } i \in \mathcal{N},$$
$$N_1 := \sum_{i\in\mathcal{N}} k(i). \tag{A.2}$$

Theorem A.1 *The $N_1 \times N_1$ symmetric nonnegative definite matrix B_p given by (6.4) decomposes into a direct sum of N' symmetric nonnegative definite matrices indexed by the set $\mathcal{N}$:*

$$B_p = \oplus_{i\in\mathcal{N}} B_{p,i}, \quad B_{p,i} = (b_{p,i}(q,r))_{q,r=1}^{k(i)}) \text{ is } k(i)\times k(i) \text{ for } i \in \mathcal{N}, \tag{A.3}$$

and

$$\mathbf{x}^{\mathrm{T}} B_p \mathbf{x} = \sum_{i\in\mathcal{N}} \mathbf{x}_i^{\mathrm{T}} B_{p,i} \mathbf{x}_i. \tag{A.4}$$

More precisely, let $\mathbf{v}_{p,k} = (v_{p,k,1},...,v_{p,k,M})^{\mathrm{T}},\ k = 1,...,M$ be given as in Algorithm 6.1. Then

$$b_{p,i}(q,r) = \sum_{k=\ell}^{M} v_{p,k,j(i,q)} v_{p,k,j(i,r)}, \quad q,r = 1,...,k(i). \tag{A.5}$$

Equivalently, let W_p be the following $M \times M$ idempotent symmetric matrix ($W_p^2 = W_p$) of rank $M - l + 1$:

$$W_p = \sum_{k=\ell}^{M} \mathbf{v}_{p,k}\mathbf{v}_{p,k}^{\mathrm{T}} = T_p T_p^{\mathrm{T}}, \quad T_p = [\mathbf{v}_{p,\ell},...,\mathbf{v}_{p,M}] \in \mathcal{M}_{M(M-\ell+1)}. \tag{A.6}$$

Then $B_{p,i}$ is the submatrix of W_p of order $k(i)$ with respect to the rows and columns in the set $\mathcal{S}(i)^c$ for $i \in \mathcal{N}$. In particular, if in each row of E there is at most one missing entry then B_p is a diagonal matrix.

Proof. View the rows and the columns of B_p as indexed by $(s,j(s,q))$ and $(t,j(t,r))$ respectively, where $s,t \in \mathcal{N}$ and $q = 1,...,k(s),\ r = 1,...,k(t)$. (For the purposes of this proof, the notation here is different from that in the body of the

paper.) So $B_p = (b_p((s, j(s,q)), (t, j(t,r))))$. Let $F(i,j)$ be the $N \times M$ matrix which has 1 on the (i,j) place and all other entries are equal to zero. Then

$$b_p((s, j(s,q)), (t, j(t,r))) =$$
$$\frac{1}{2}\sum_{i=\ell}^{M} \mathbf{v}_{p,i}^{\mathrm{T}}(F(s, j(s,q))^{\mathrm{T}}F(t, j(t,r)) + F(t, j(t,r))^{\mathrm{T}}F(s, j(s,q)))\mathbf{v}_{p,i}, \quad \text{(A.7)}$$
$$s, t \in \mathcal{N},\ q = 1, ..., k(s), r = 1, ..., k(t).$$

It is straightforward to show that $F(s, j(s,q))^{\mathrm{T}}F(t, j(t,r)) = 0$ if $s \neq t$. Furthermore, for $s = t$ the matrix $F(s, j(s,q))^{\mathrm{T}}F(t, j(t,r)) + F(t, j(t,r))^{\mathrm{T}}F(s, j(s,q))$ has 1 in the places $(j(s,q), j(t,r))$ and $(j(t,r), j(s,q))$ for $r \neq q$, and has 2 in the place $(j(s,q), j(s,q))$ if $r = q$ and zero in all other positions. Hence
$b_p((s, j(s,q)), (t, j(t,q))) = 0$ unless $s = t$. If $s = t$ then a straightforward calculation yields (A.5). Other claims of the theorem follow straightforward from the equality (A.5). □

B Algorithm for 6.7

From Theorem A.1, the system of equations $B_p\mathbf{x} = -\mathbf{w}_p$ in N_1 unknowns is equivalent to N' smaller systems

$$B_{p,i}\mathbf{x}_{p+1,i} = -\mathbf{w}_{p,i} \quad i \in \mathcal{N}. \quad \text{(B.1)}$$

Thus the big system of equations in N_1 unknowns, the coordinates of $\mathbf{x}_{p+1}$, given (6.7) splits to N' independent systems given in (B.1). That is, in the iterative update of the unknown entries of E given by the matrix E_{p+1}, the values in the row $i \in \mathcal{N}$ in the places $\mathcal{S}(i)^c$ are determined by the values of the entries of E_p in the places $\mathcal{S}(i)^c$ and the eigenvectors $\mathbf{v}_{p,\ell}, ..., \mathbf{v}_{p,M}$ of $E_p^{\mathrm{T}}E_p$.

We now show how to efficiently solve the system (6.7).

Algorithm B.1 *For $i \in \mathcal{N}$ let $V_{p,i}$ is the $k(i) \times (M - \ell + 1)$ matrix obtained from V_p, given by (A.6), by deleting all rows except the rows $j(i,1), ..., j(i,k(i))$. Then (B.1) is equivalent to*

$$V_{p,i}V_{p,i}^{\mathrm{T}}\mathbf{x}_{p+1,i} = -\mathbf{w}_{p,i}, \quad i \in \mathcal{N}, \quad \text{(B.2)}$$

which can be solved efficiently by the QR algorithm as follows. Write $V_{p,i}$ as $Q_{p,i}R_{p,i}P_{p,i}$, where $Q_{p,i}$ is an $k(i) \times d_{p,i}$ matrix with $d_{p,i}$ orthonormal columns, $R_{p,i}$ is an upper triangular $d_{p,i} \times k(i)$ matrix of rank $d_{p,i}$ nonzero rows, where the rank $V_{p,i} = d_{p,i}$, and $P_{p,i}$ is a permutation matrix. (The columns of $Q_{p,i}$ are obtained from the columns of $V_{p,i}$ using Gram-Schmidt process.) Then

$$Q_{p,i}^{\mathrm{T}}\mathbf{x}_{p+1,i} = -(R_{p,i}R_{p,i}^{\mathrm{T}})^{-1}Q_{p,i}^{\mathrm{T}}w_{p,i}$$

and

$$\mathbf{x}_{p+1,i} = -Q_{p,i}(R_{p,i}R_{p,i}^{\mathrm{T}})^{-1}Q_{p,i}^{\mathrm{T}}w_{p,i}, \quad i \in \mathcal{N} \tag{B.3}$$

is the least square solution for $\mathbf{x}_{p+1,i}$.

C Matlab code

```
function Ep1 = fraa(E,Ep,L,iter)
%Fixed rank algorithm -- estimate missing values
%Usage: fraa(E,Ep,L,iter)
%E: matrix with missing values
%Ep: initial solution
%L: parameter (number of significant values)
%iter: number of iterations to perform
%Note: Any rows with all missing values must be removed
%%%%%%%%%% THIS IS  THE SET-UP
%Get size of E
    [N,M]=size(E);
       if (L > M)
        error('L must be less than or equal to the number of columns of E')
      end;
  %get index of missing values
   missing=find(isnan(E));
   %Number of missing values
    m=length(missing);
    m2=m*m;
%%%%%%%%%% NOW WE WORK WITH THE ALGORITHM
  Xp1=zeros(N,M);
   track=iter;
 while(iter > 0)
   A=Ep'*Ep;
   %Find singular value decomposition of A
   [U,S,V]=svd(A);
   %Singular values of Ep
   sigma2=S(S~=0);
   singular=sqrt(sigma2);
   partial_sig2=sum(sigma2(L:M));
   total_sig2=sum(sigma2(1:M));
   fprintf('\n iteration %3.0f \n', track-iter+1)
   fraction=partial_sig2/total_sig2;
```

```
   fprintf(' partial sum/total sum of sq. singular values \n %1.8f', fraction)
   fprintf('\n')
%Construct B=Bp
     B=sparse(m,m);  %pre-allocate space
      [is,js]=ind2sub([N,M],missing(1:m));
       for s=1:m
         for t=s:m
            if (i(s)==i(t))
            B(s,t)=sum(U(j(s),L:M)*U(j(t),L:M)');
            B(t,s)=B(s,t);    %B is symmetric
           end  %end if
        end %end For t
      end %end for s
%%%NOW CONSTRUCT THE VECTOR Wp
W=sparse(m,1); %pre-allocate space
    for t=1:m
     K=sparse(N,M);
     K(missing(t))=1;
     W(t)=sum(diag(U(:,L:M)'*Ep'*K*U(:,L:M)));
   end %end for
  %Solve Bx_(p+1)= -W
   xp1=-B\W;
  %Create matrix B_{p+1}
   Xp1(missing)=xp1;
  %Update solution
   Ep=Ep+Xp1;
  %set counter
   iter=iter-1;
  end %End while
   fprintf('\n')
   fprintf(' singular values (final iteration):\n')
   fprintf('%16.6f',singular)
   Ep1=Ep;
```

For the Matlab m file or a version of this algorithm for R, see http://www.carleton.edu/~faculty/lchihara

References

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

[2] 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.

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

[4] 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.

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

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

[7] 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.

[8] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R. Altman, Missing value estimatation for DNA microarrays, Bioinformatics 17 (2001), 1-6.

Figure captions:

7.1 Comparison of normalized RMS against percent missing for threee methods: FRAA, KNNimpute, and row means methods. The normalized RMS for the zeros method is not displayed, but the values are $.397, .870, 1.24, 1.52, 1.76$, for $1, 5, 10, 15, 20\%$ percent missing, respectively.

7.2 Four columns of the full elutriation matrix. Entries were then randomly deleted. (a) Plot of normalized RMS against percent missing. (b) Distribution of the raw errors (true - estimate) in one run of a simulation with 10% missing.

7.3 Scatter plot of the raw errors from the KNNimpute and FRAA estimates of the truncated elutriation matrix with 10% entries missing. The correlation between the two sets of raw errors is .76.

Fig. 7.1

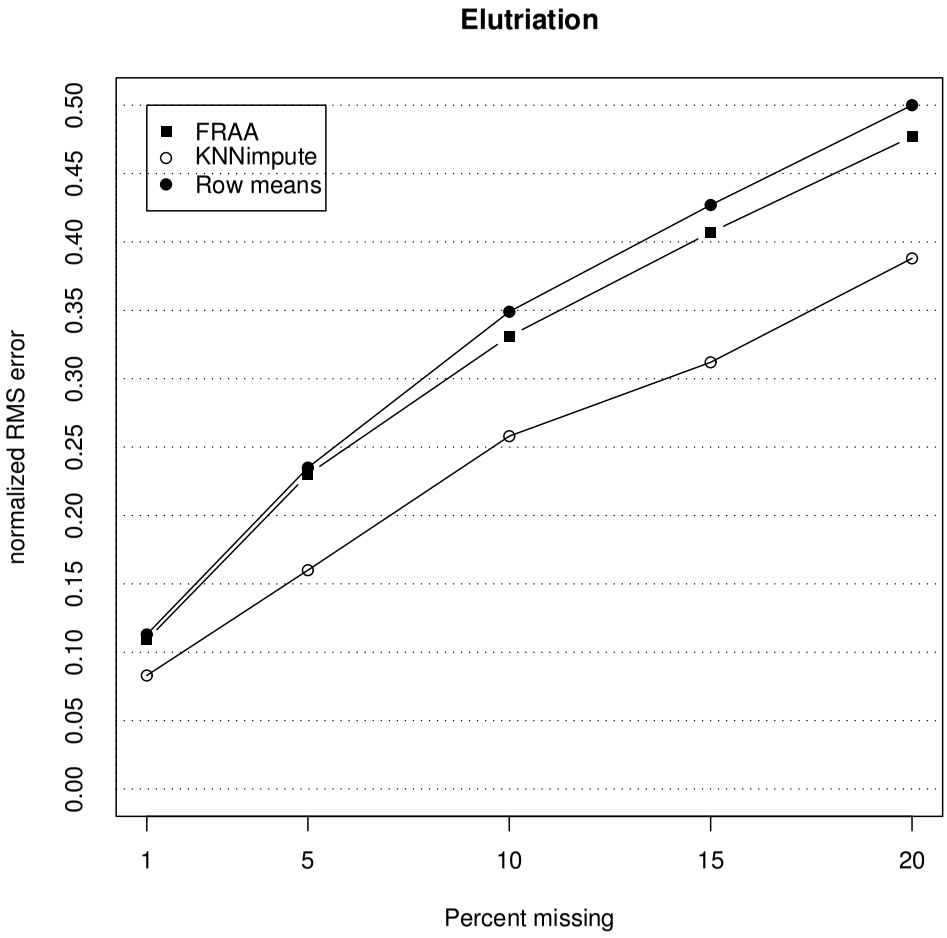

Elutriation
FRAA
KNNimpute
Row means
normalized RMS error
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
1
5
10
15
20
Percent missing

Fig. 7.2(a)

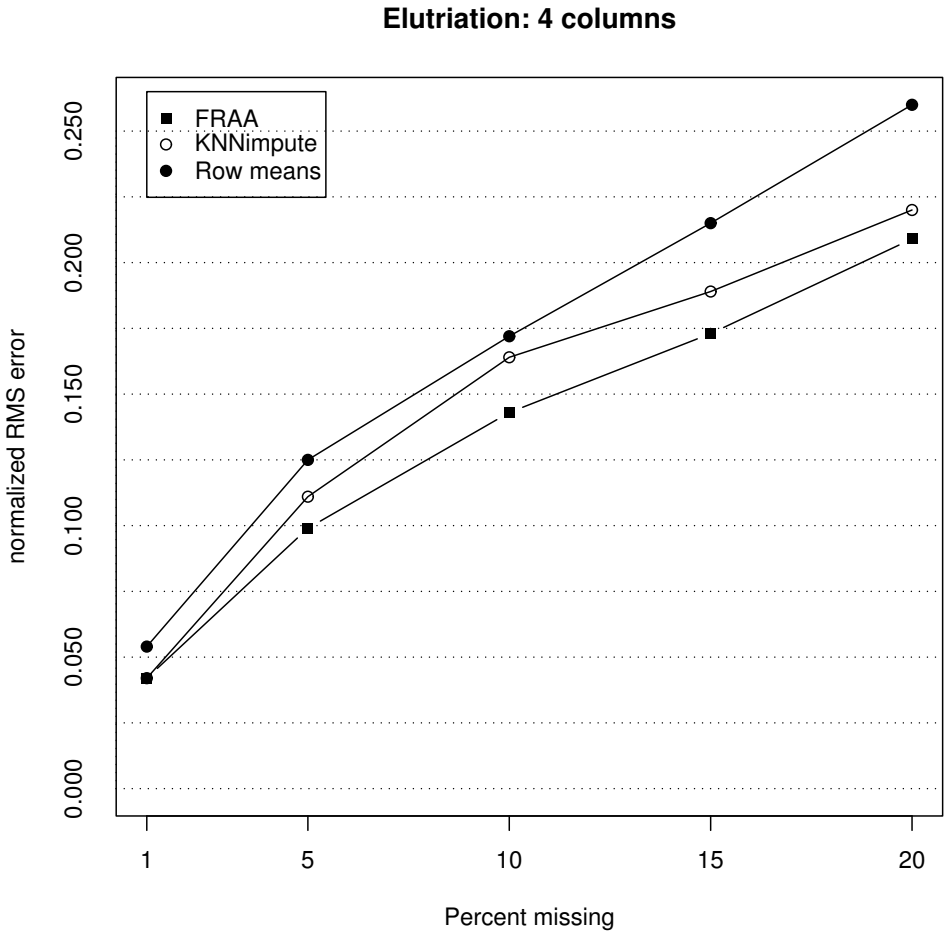

Elutriation: 4 columns
FRAA
KNNimpute
Row means
normalized RMS error
0.000
0.050
0.100
0.150
0.200
0.250
1
5
10
15
20
Percent missing

Fig. 7.2(b)

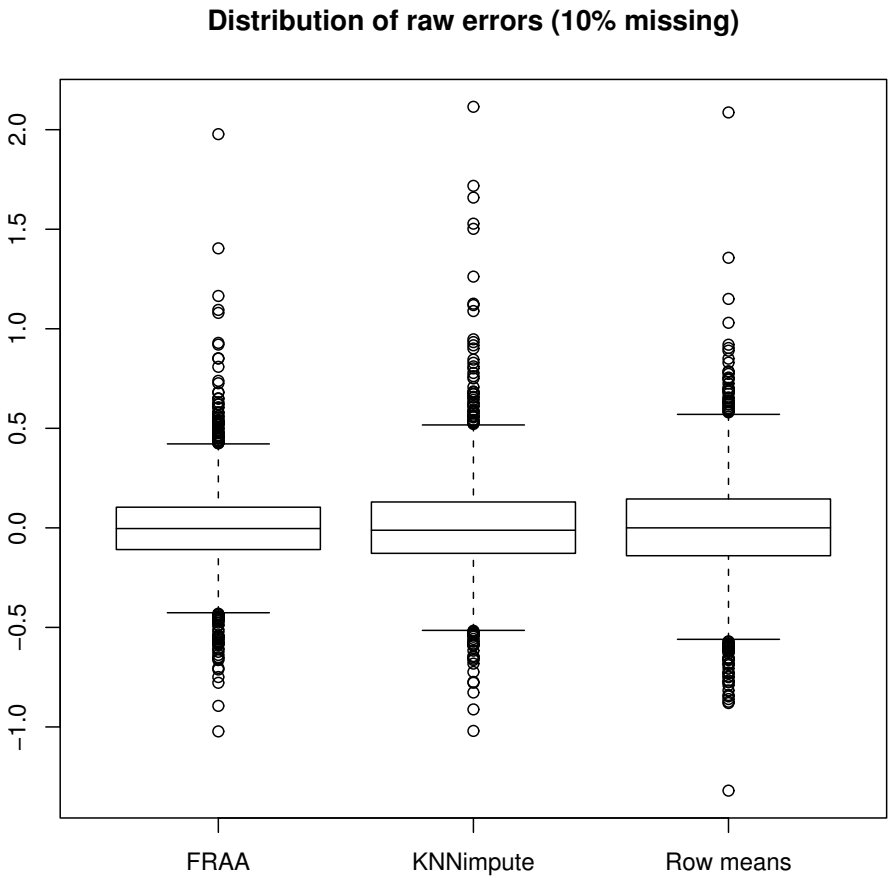
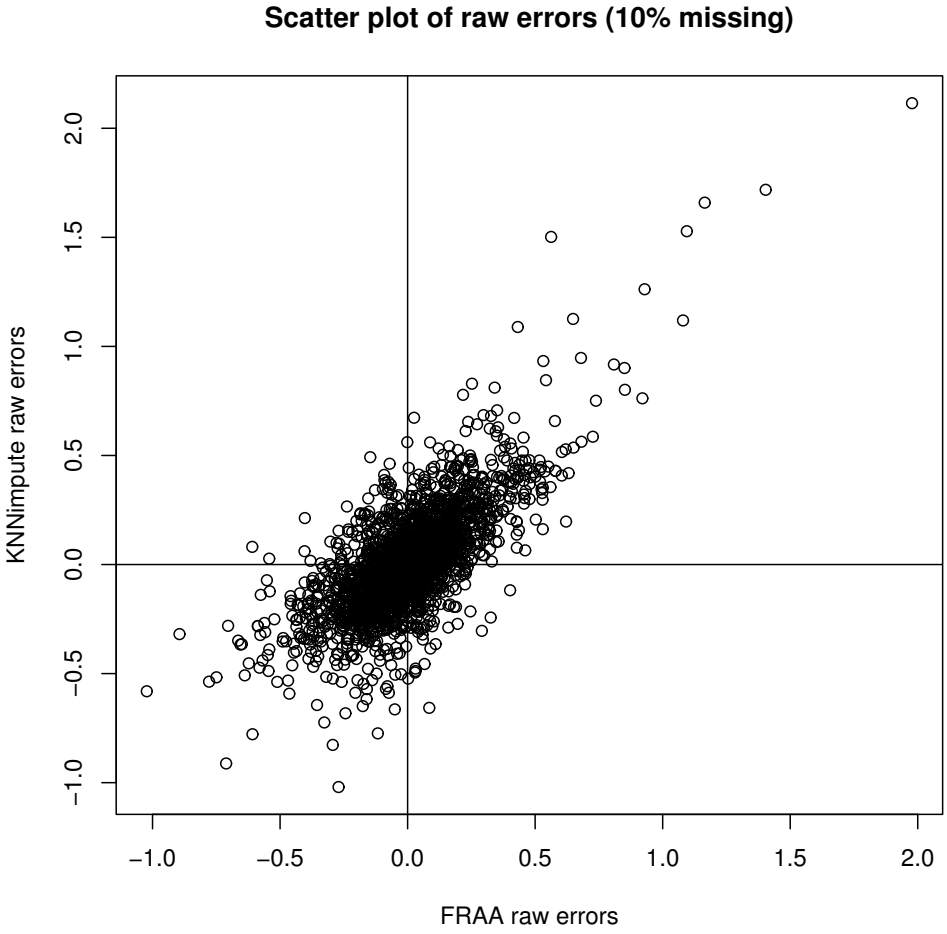

Distribution of raw errors (10% missing)
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
FRAA
KNNimpute
Row means

Fig. 7.3


Scatter plot of raw errors (10% missing)
KNNimpute raw errors
FRAA raw errors
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0