Statistical Applications in Genetics and Molecular Biology

Volume 4, *Issue* 1 2005 *Article* 10

Prediction of Missing Values in Microarray and Use of Mixed Models to Evaluate the Predictors

Guri Feten* Trygve Almøy†

Are H. Aastveit‡

*Norwegian University of Life Sciences, guri.feten@umb.no

†Norwegian University of Life Sciences, trygve.almoy@umb.no

‡Norwegian University of Life Sciences, are.aastveit@umb.no

Prediction of Missing Values in Microarray and Use of Mixed Models to Evaluate the Predictors*

Guri Feten, Trygve Almøy, and Are H. Aastveit

Abstract

Gene expression microarray experiments generate data sets with multiple missing expression values. In some cases, analysis of gene expression requires a complete matrix as input. Either genes with missing values can be removed, or the missing values can be replaced using prediction. We propose six imputation methods. A comparative study of the methods was performed on data from mice and data from the bacterium *Enterococcus faecalis*, and a linear mixed model was used to test for differences between the methods. The study showed that different methods' capability to predict is dependent on the data, hence the ideal choice of method and number of components are different for each data set. For data with correlation structure methods based on K-nearest neighbours seemed to be best, while for data without correlation structure using the average of the gene was to be preferred.

*We would like to thank two anonymous referees for constructive criticism. This work has been financed by Norwegian University of Life Sciences.

1 Introduction

The technology of DNA microarray allows monitoring expression levels for thousands of genes simultaneously. Large-scale gene expression studies have been carried out to study cell cycle (Eisen *et al.*, 1998), tumor tissues (DeRisi *et al.*, 1996), yeast sporulation (Chu *et al.*, 1998), and resequence and mutational analysis (Hacia, 1999). An introduction to the microarray technology can be found in e.g. Nguyen *et al.* (2002a).

The microarray data are characterized by many measured variables (genes) on only a few observations (replications, parallels). Often microarray experiments generate data sets with multiple missing expression values. In some cases, analysis of gene expression requires a complete matrix as input, e.g. hierarchical clustering.

There are different reasons for missing expression values. The microarray may contain weak spots. Usually these spots are filtered out. One way to sort out weak spots is to compare the pixels of the spot with the pixels of the background. If the fraction of spot pixels greater than the median of the background pixels is less than a given threshold, the gene expression that corresponds to this spot will be set as missing. Another reason for missing expression values is technical errors during the hybridization. Microarrays are scanned in a microarray scanner, producing either fluorescence intensities or radioactive intensities. The intensities must be higher than a given value. If the intensity is below this threshold, we define the value as missing. A third reason for missing values is dust, scratches, and systematic errors on the slides.

Recently comparative studies of three data imputation methods; a singular value decomposition based method, weighted K-nearest neighbours, and row average were presented (Troyanskaya *et al.*, 2001; Hastie *et al.*, 1999). Bø *et al.* (2004) compared methods that utilize correlations between both genes and arrays based on the least square principle and the method of K-nearest neighbours. Ouyang *et al.* (2004) proposed an imputation method based on Gaussian mixture clustering and model averaging. All of these papers investigated the methods for different fractions of missing data. Nguyen *et al.* (2004) investigated how the accuracy of four different prediction methods; mean, K-nearest neighbours, ordinary lest squares regression, and partial least squares regression, is depended on the actual gene expression. In addition to the work that has been done on missing value prediction for microarray data, larger studies have been devoted to similar problems in other fields. Common methods include iterating procedures (Yates, 1933; Healy and Westmacott, 1956), imputing conditional means (Buck, 1960), hot deck imputation (Ernst, 1980), multiple imputation (Rubin, 1987; Rubin and Schenker, 1991), and bootstrap (Efron, 1994).

In this paper different methods for replacing missing data have been studied. In addition to previously studied methods; Principal Component Regression (PCR), Partial Least Square Regression (PLSR), weighted

K-nearest neighbours (KNN) with genes as neighbours, and gene average, we focused on two methods; Factor Analysis Regression (FAR) and weighted K-nearest neighbours (KNN) with observations as neighbours.

Earlier papers on missing values have none or few attempts on comparing the prediction methods statistically. Bø *et al.* (2004) used paired t-test to compare the methods. Cross Validation Analysis of Variance (CVANOVA) was introduced to compare prediction methods by Indahl and Næs (1998). Based on the idea of CVANOVA, in this paper we have used mixed models to compare the methods.

In Section 2 we will present six methods to predict missing values, four based on regression, and two based on K-nearest neighbours. A method based on linear mixed models to compare the prediction methods is also described. Further on, in Section 3, there is a presentation of the data used in the study, followed, in Section 4, by the results of the study.

2 Methods

2.1 Prediction and missing values

Suppose we will study the gene expressions of p genes (typically $1000 - 40000$) on n observations ($p \gg n$), where several of the genes have missing expressions. Our aim is to predict these missing values. Let $\mathbf{y}_j$ denote the $n \times 1$ vector containing the n gene expressions of gene j. Let $\mathbf{X}_{(j)}$ denote the $n \times (p-1)$ matrix containing the gene expressions of the $p-1$ other genes. If nothing else is stated, both $\mathbf{y}_j$ and $\mathbf{X}_{(j)}$ are centered by subtracting their column averages.

Some methods for predicting missing values are based on the regression approach. Since p is larger than n, common methods as least square or maximum likelihood does not apply. Hence other prediction methods have to be used for the purpose of reducing the predictor space spanned by the $(p-1)$ columns in $\mathbf{X}_{(j)}$ to a lower K-dimensional space. This is achieved by constructing K components in the predictor space, where the components optimize a defined object criterion.

If observation i has a missing value for gene j, it can generally be predicted by a linear predictor $\hat{y}_{ij}$ given by

$$\hat{y}_{ij} = \bar{y}_j + \hat{\boldsymbol{\beta}}_j^t(\mathbf{x}_{(j)i} - \bar{\mathbf{x}}_{(j)}), \tag{1}$$

where $\bar{y}_j$ is the average of the uncentered expression of gene j, $\mathbf{x}_{(j)i}$ is a vector of the expressions of all the other genes of the observation with the missing value, and $\bar{\mathbf{x}}_{(j)}$ is a vector containing the average of the uncentered gene expression for each gene except gene j. To simplify the notation we let $\mathbf{y} = \mathbf{y}_j$, $\mathbf{X} = \mathbf{X}_{(j)}$, and $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_j$ in the rest of the paper. We will in the following present different methods for computing $\hat{\boldsymbol{\beta}}$, methods that are

known from other fields to perform well, e.g. chemometrics (Martens and Næs, 1989).

The methods predict the missing values by aim of an iterative method based on the idea of the Expectation Maximization (EM) algorithm (Dempster *et al.*, 1977; Wu, 1983). For iteration q, define the matrix $\mathbf{W}^{(q)} = \left[\begin{array}{cc} \mathbf{y}^{(q)} & \mathbf{X}^{(q)} \end{array} \right]$. Since methods based on regression can only be performed on complete matrices, we initially (that is $q = 0$) set the missing values equal to the average of the non-missing values for each gene. Let $q = q + 1$, and then for each gene with missing values, compute the vector $\hat{\boldsymbol{\beta}}$ in (1) from $\mathbf{y}^{(q-1)}$ and $\mathbf{X}^{(q-1)}$. Furthermore, $\mathbf{y}^{(q)}$, is produced by replacing the missing values in $\mathbf{y}$ with the fitted values from (1), and then update $\mathbf{W}^{(q)}$. The computation is repeated until $\|\mathbf{W}^{(q-1)} - \mathbf{W}^{(q)}\|/\|\mathbf{W}^{(q-1)}\|$ is below some threshold, e.g. 10^{-2}, where $\|\mathbf{W}\| = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{p} w_{ij}^2}$ (Hastie *et al.* 1999).

2.2 Prediction methods

2.2.1 Parametric methods

In situations where there are more variables than observations, the matrix $\mathbf{X}^t\mathbf{X}$ will be singular, and hence $(\mathbf{X}^t\mathbf{X})^{-1}$ does not exist and ordinary least square cannot be used. Hopefully a few linear combinations of the variables will take care of most of the available information in data, and the remaining combinations could be declared as noise and then removed.

The regression methods studied in this paper have in common that they express the $(p-1)$-dimensional vector $\hat{\boldsymbol{\beta}}$ in (1) by

$$\hat{\boldsymbol{\beta}} = \mathbf{R}(\mathbf{R}^t\mathbf{S}\mathbf{R})^{-1}\mathbf{R}^t\mathbf{s}, \tag{2}$$

where $\mathbf{S} = \mathbf{X}^t\mathbf{X}$, $\mathbf{s} = \mathbf{X}^t\mathbf{y}$, and the $p \times K$ matrix $\mathbf{R}$ are specified by the regression methods. The vector $\hat{\boldsymbol{\beta}}$ in (2) can be written as

$$\hat{\boldsymbol{\beta}} = \sum_{k=1}^{K} a_k \mathbf{b}_k. \tag{3}$$

Geometrically $\hat{\boldsymbol{\beta}}$ is element of a K-dimensional ($K \ll p$) subspace spanned by the vectors $\mathbf{b}_1, \ldots, \mathbf{b}_K$. The constants a_k and the $p \times 1$ vectors $\mathbf{b}_k$ are specified by the regression methods. The optimal number of components (K) for getting the best prediction, needs to be determined empirically. The topic is discussed by Næs and Helland (1993) and Helland and Almøy (1994).

The average The simplest method to predict missing values is to use the average over the uncentered expression values for the associated gene. This is equal to assuming $\hat{\boldsymbol{\beta}} = \mathbf{0}$. The predictor in (1) is then simplified to

$$\hat{y}_{ij} = \bar{y}_j.$$

We call the method "average".

Principal Component Regression (PCR) In Principal Component Analysis (PCA), we will explain as much of the variance-covariance structure as possible through a few linear combinations of the original explanatory variables (Massy, 1965). The PCR method is equivalent to applying these combinations as explanatory variables, and performing a least square regression on the new variables. This transformation eliminates the collinearity between the variables, and the stability of the regression coefficients is increased, while some bias is introduced. The scores of the principal components are given by

$$\mathbf{Z} = \mathbf{XE},$$

where $\mathbf{E}$ is the orthogonal matrix whose K columns are the eigenvectors of $\mathbf{X}^t\mathbf{X}$. The eigenvectors' contribution to the expression is quantified by the corresponding eigenvalues. We identify the most significant eigenvectors by sorting them based on their corresponding eigenvalues. The components having small eigenvalues are assumed corresponding to noise.

We get $\hat{\boldsymbol{\beta}}$ with PCR if we insert

$$\mathbf{R} = \mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_K]$$

in equation (2), or equivalently insert

$$a_k = \frac{\mathbf{e}_k^t \mathbf{s}}{l_k} \text{ and } \mathbf{b}_k = \mathbf{e}_k$$

in (3), where l_k is the eigenvalue corresponding to the eigenvector $\mathbf{e}_k$. In PCR we therefore use fewer variables, but keep the important information.

Partial Least Square Regression (PLSR) This method is well known from chemometrics (Martens and Næs, 1989), and has previously been used in classification based on microarray gene expression data (Nguyen and Rocke, 2002b; Ghosh, 2003). It is shown that PLSR is useful as a predictive modelling regression method in the kind of data where there are many more variables than observations (Næs *et al.*, 1986; Höskuldsson, 1988).

We receive $\hat{\boldsymbol{\beta}}$ with PLSR by inserting

$$\mathbf{R} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_K]$$

in (2), where the vector $\mathbf{s}_k = \mathbf{X}_k^t \mathbf{y}_k$ is the non-normalized eigenvector corresponding to the largest eigenvalue of $\mathbf{X}_k^t \mathbf{y}_k \mathbf{y}_k^t \mathbf{X}_k$, and where $\mathbf{X}_k$ and $\mathbf{y}_k$ denote the residual after k components, given by

$$\begin{aligned} \mathbf{y}_k &= \mathbf{y}_{k-1} - \mathbf{X}_{k-1}\mathbf{s}_{k-1}(\mathbf{s}_{k-1}^t \mathbf{S}_{k-1}\mathbf{s}_{k-1})^{-1}\mathbf{s}_{k-1}^t \mathbf{s}_{k-1}, \\ \mathbf{X}_k &= \mathbf{X}_{k-1} - \mathbf{X}_{k-1}\mathbf{s}_{k-1}(\mathbf{s}_{k-1}^t \mathbf{S}_{k-1}\mathbf{s}_{k-1})^{-1}\mathbf{s}_{k-1}^t \mathbf{S}_{k-1}, \end{aligned}$$

where $\mathbf{S}_{k-1} = \mathbf{X}_{k-1}^t\mathbf{X}_{k-1}$, $\mathbf{X}_1 = \mathbf{X}$, and $\mathbf{y}_1 = \mathbf{y}$. There is no simple expression of the a_k's and $\mathbf{b}_k$'s in (3), the exception is one component, then

$$a_1 = \frac{\mathbf{s}_1^t\mathbf{s}_1}{\mathbf{s}_1\mathbf{S}_1\mathbf{s}_1} \text{ and } \mathbf{b}_1 = \mathbf{s}_1.$$

Factor Analysis Regression (FAR) In contrast to PCR and PLSR, the FAR method is based on a statistical model. Let y be the expression of a gene with missing values, and let $\mathbf{x}$ be a vector containing the expressions of the $(p-1)$ other genes. Let us assume that both $\mathbf{x}$ and y simultaneously follow the factor analysis model given by (Lawley and Maxwell, 1973)

$$\begin{bmatrix} y \\ \mathbf{x} \end{bmatrix} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\mathbf{f} + \boldsymbol{\epsilon}, \tag{4}$$

where $\boldsymbol{\mu}$ $(p \times 1)$ is a vector of constants, $\boldsymbol{\Gamma}$ $(p \times r)$ is a matrix of factor loadings, $\mathbf{f}$ $(r \times 1)$ is a vector of factor scores, and $\boldsymbol{\epsilon}$ $(p \times 1)$ is a vector of specific factors with equal variance. Let $\mathbf{f} \sim N(\mathbf{0}, \mathbf{I}_r)$, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \psi\mathbf{I}_p)$, and $Cov(\mathbf{f},\boldsymbol{\epsilon}) = \mathbf{0}$ $(r \times p)$, then the covariance matrix of $\begin{bmatrix} y & \mathbf{x}^t \end{bmatrix}^t$ is given by

$$\boldsymbol{\Gamma}\boldsymbol{\Gamma}^t + \psi\mathbf{I}_{p+1} = \begin{bmatrix} \boldsymbol{\gamma}_y^t\boldsymbol{\gamma}_y + \psi & \boldsymbol{\gamma}_y^t\boldsymbol{\Gamma}_\mathbf{x}^t \\ \boldsymbol{\Gamma}_\mathbf{x}\boldsymbol{\gamma}_y & \boldsymbol{\Gamma}_\mathbf{x}\boldsymbol{\Gamma}_\mathbf{x}^t + \psi\mathbf{I}_p \end{bmatrix}. \tag{5}$$

More details regarding the factor model can be found in any textbook in multivariate analysis, e.g. Mardia *et al.* (1979).

If we assume that $\boldsymbol{\Gamma}$ is a matrix with orthogonal columns, then cumbersome, but straightforward calculation gives the following maximum likelihood estimators of the parameters in (5)

$$\begin{aligned} \hat{\psi} &= (p-r)^{-1}\sum_{k=r+1}^{p} l_k = \bar{l}_r, \\ \hat{\boldsymbol{\Gamma}}_\mathbf{x} &= \mathbf{H}_r(\mathbf{L}_r - \bar{l}_r\mathbf{I}_r)^{0.5}, \\ \hat{\boldsymbol{\gamma}}_y &= (\mathbf{L}_r - \bar{l}_r\mathbf{I}_r)^{0.5}\mathbf{g}_r. \end{aligned}$$

Here l_k is the k'th largest eigenvalues of $([\mathbf{y},\mathbf{X}]^t[\mathbf{y},\mathbf{X}])$, $\mathbf{L}_r$ is the diagonal matrix with the r largest eigenvalues, and $\begin{bmatrix} \mathbf{g}_r, & \mathbf{H}_r^t \end{bmatrix}^t$ is the $p \times r$ matrix of the corresponding eigenvectors (Almøy, 1994). Assuming (5) we obtain by straightforward matrix calculation the following maximum likelihood regression vector for

$$\boldsymbol{\beta} = (\boldsymbol{\Gamma}_\mathbf{x}\boldsymbol{\Gamma}_\mathbf{x}^t + \psi\mathbf{I})^{-1}\boldsymbol{\Gamma}_\mathbf{x}\boldsymbol{\gamma}_y = \boldsymbol{\Gamma}_\mathbf{x}(\boldsymbol{\Gamma}_\mathbf{x}^t\boldsymbol{\Gamma}_\mathbf{x} + \psi\mathbf{I})^{-1}\boldsymbol{\gamma}_y$$

as

$$\hat{\boldsymbol{\beta}} = \mathbf{H}_r(\mathbf{L}_r - \bar{l}_r\mathbf{I}_r)\mathbf{L}_r^{-1}\mathbf{g}_r[1 - \mathbf{g}_r^t(\mathbf{L}_r - \bar{l}_r\mathbf{I}_r)\mathbf{L}_r^{-1}\mathbf{g}_r]^{-1},$$

or expressed as in (3) where

$$a_k = g_k(1 - \bar{l}_K l_k^{-1})(1 - \mathbf{g}_K^t(\mathbf{L}_K - \bar{l}_K\mathbf{I}_K)\mathbf{L}_K^{-1}\mathbf{g}_K)^{-1} \text{ and } \mathbf{b}_k = \mathbf{h}_k,$$

with g_k being the k'th element of $\mathbf{g}_K$, and $\mathbf{h}_k$ being the k'th vector of $\mathbf{H}_K$.

2.2.2 Non-parametric methods

K-Nearest Neighbours (KNN) While all the previously presented methods are parametric, another way to predict missing values is by the application of some non-parametric methods. The most used method is K-nearest neighbours, which predict missing values by imputing them with a weighted average of their neighbours (Troyanskaya *et al.*, 2001). The neighbours can be either observations or genes. We assume that there exist one or several groups of coregulated genes, hence we can use genes as neighbours.

To define a neighbour we need a distance measure between the object with the missing values of interest and the neighbours. In this paper we have used Euclidean distance based on all pairs of non-missing data. Different objects have different numbers of complete pairs, hence the sum is scaled up proportionally to the number of complete pairs. When observations are neighbours use $\mathbf{X}' = \mathbf{X}$, and when genes are neighbours use $\mathbf{X}' = \mathbf{X}^t$. The distance between two rows (i and l) is given by

$$d_{il} = \sqrt{\frac{p}{c_{il}} \sum_{j \in C_{il}} (x'_{ij} - x'_{lj})^2}, \tag{6}$$

where p is the number of columns in $\mathbf{X}'$, c_{il} is the number of complete pairs of rows i and l, C_{il} is the set of complete pairs of rows i and l, x'_{ij} is the value in column j and row i, and x'_{lj} is the value in column j and row l. The algorithm of K-nearest neighbours uses imputed values from objects close to the object with the missing value (Little and Rubin, 1987). We have used an extended version where neighbours are weighted down according to increasing distance (Troyanskaya *et al.*, 2001).

Consider a row i that has one missing value in column j, the KNN method finds K other rows which have a value present in column j, with value most similar to row i. The missing value in column j and row i can then be replaced by using

$$\hat{x}'_{ij} = \bar{x}'_{(i)} + \frac{\sum_{l \neq i}(x'_{lj}/d_{il})}{\sum_{l \neq i}(1/d_{il})},$$

where $\bar{x}'_{(i)}$ is the average value of the uncentered row i, x'_{lj} is the l'th value in column j, and d_{il} is the Euclidean distance given in (6) (Little and Rubin, 1987). Note that when we use genes as neighbours, we center the observations instead of the variables.

2.3 Validation

The original data set was pre-processed by removing the genes containing missing expression values, yielding complete matrices. To create test data sets we deleted randomly between 0.1% and 20% of the data from the

complete matrix. Each method was then used to recover the introduced missing values for the data set. The predicted values were compared to those in the original data set. We determined the method with optimal number of components for every fraction of data missing.

For Principal Component Regression, Partial Least Square Regression, and Factor Analysis Regression different numbers of components were used. For K-nearest neighbours, both the number of neighbouring genes and the number of neighbouring observations optimal for prediction were varied.

The ability of the predictor is usually evaluated by its expected square loss,

$$\theta^2_{jk(m)} = E_j(y_j - \hat{y}_{jk(m)})^2,$$

interpreted as the long run average square difference between the uncentered expression of gene j and the predicted expression using k components within method m. Since θ is a complicated function of unknown parameters, it can be estimated by the root mean square error of prediction (RMSEP). The commonly used RMSEP is given by

$$RMSEP = \sqrt{\frac{1}{\sum_{j\in Q} t_j} \sum_{j\in Q} \sum_{i\in T_j} (y_{ij} - \hat{y}_{ijk(m)})^2},$$

where Q is the set of genes with missing values, t_j is the number of missing values in gene j, T_j is the set of observations with missing values in gene j. Further on, y_{ij} is the observed uncentered expression of observation i in gene j, and $\hat{y}_{ijk(m)}$ is the predicted value of observation i in gene j using method m with k components/neighbours. Only if all the genes, or the number of missing values is equal for all genes, $RMSEP^2$ is an unbiased estimate of the average (over genes) prediction error,

$$\theta^2_{k(m)} = \frac{1}{q} \sum_{j\in Q} \theta^2_{jk(m)}, \tag{7}$$

where q is the number of genes with missing values. Instead we have used a modified version of RMSEP, whose square value is an unbiased estimate of $\theta^2_{k(m)}$, given by

$$RMSEP_{k(m)} = \sqrt{\frac{1}{q} \sum_{j\in Q} \left(\frac{1}{t_j} \sum_{i\in T_j} (y_{ij} - \hat{y}_{ijk(m)})^2 \right)}. \tag{8}$$

Comparison of methods In order to compare the different methods for predicting missing values, the predictors may be compared using ideas from CVANOVA (Cross Validation Analysis of Variance) introduced by Indahl and Næs (1998). This method is based on two-way analysis of variance of prediction results obtained using cross-validation.

Here we will generalize this method and base the analyses on the mixed model

$$(y_{ij} - \hat{y}_{ijk(m)})^2 = \mu + \tau_i + \eta_j + \xi_m + \delta_{k(m)} + (\eta\xi)_{jm} + (\eta\delta)_{jk(m)} + \epsilon_{ijk(m)}, \quad (9)$$

where μ is the common average, τ_i is the effect of the i'th observation, η_j is the effect of the j'th gene, ξ_m is a parameter associated with the m'th method, $\delta_{k(m)}$ is the effect of the k'th component within method m, $(\eta\xi)_{jm}$ is the effect of the interaction between the j'th gene and the m'th method. Further on, $(\eta\delta)_{jk(m)}$ is the effect of the interaction between the j'th gene and the k'th component within method m, and $\epsilon_{ijk(m)}$ is a random error component. In the model we will assume $\eta_j \sim N(0, \sigma^2_\eta)$, $(\eta\xi)_{jm} \sim N(0, \sigma^2_{\eta\xi})$, and $\epsilon_{ijk(m)} \sim N(0, \sigma^2_\epsilon)$. Since we will only apply our conclusions to the observations considered in the analysis, the effect of observation is fixed.

The data were analyzed as a mixed model, and the hypotheses that there is effect of method, component(method), gene, gene*method, and gene*component(method) were tested by PROC GLM in SAS (The SAS System for Windows, release 8.02, SAS Institute Inc). These tests are usually robust against deviations from the normal distribution (Lindman, 1992), and the problem of using the non-normal $(y_{ij} - \hat{y}_{ijk(m)})^2$ should be minimal. We could have used the absolute error of prediction, $|y_{ij} - \hat{y}_{ijk(m)}|$, but since we evaluate the methods by RMSEP, and RMSEP uses the square error of prediction, we do not.

3 Data

To demonstrate the different methods for predicting missing values we have used data from two cDNA microarray experiments as examples.

The first is gene expression data from a study of lipid metabolism in mice focusing on identifying genes with altered expression. A mouse model with very low High Density Lipoprotein (HDL) cholesterol levels was compared to inbred control mouse. The treatment group consisted of 8 mice where the apolipoprotein (apo) AI gene was knocked out. Apo AI is a gene playing an important role in the HDL metabolism. The control group consisted of 8 inbred "normal" mice. Gene expressions from 5548 genes were measured, including 200 related to lipid metabolism. We removed all genes containing missing values, yielding 5486 genes. We denote this data set apo AI. (Callow *et al.*, 2000.) The eigenvalues of the covariance matrix of apo AI are presented in Figure 1a. The fact that one of the eigenvalues is large compared to the others is an indication of strong correlation structure between the genes.

The second example is gene expression data from a study of bacterial response to stress (not published earlier). The reference samples were normal cells of the bacterium *Enterococcus faecalis* (denoted V583), and the test samples were cells from the same bacterium, but stressed with nisin.

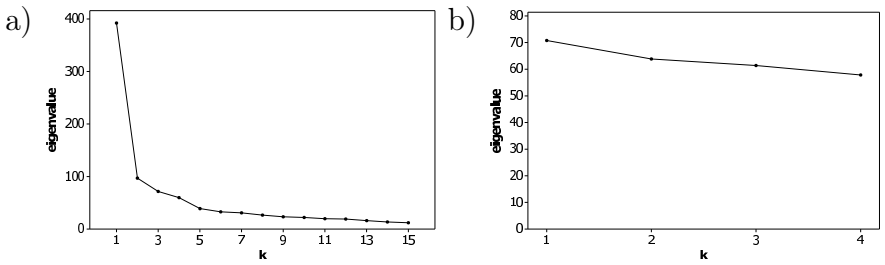

Figure 1: The eigenvalues of apo AI (a) and V583 (b).

There were five replicates of every gene on the array. Gene expressions from 3245 genes were measured. We removed all genes containing missing values, yielding 2744 genes. The eigenvalues of the covariance matrix of V583 are presented in Figure 1b. All the eigenvalues are approximately equal. This is an indication of weak correlation structure between the genes.

4 Results

We are interested in knowing whether the results of the methods are dependent of the fraction of missing values. Figure 2 presents the minimum RMSEP for each method at different fractions of missing values. Notice that

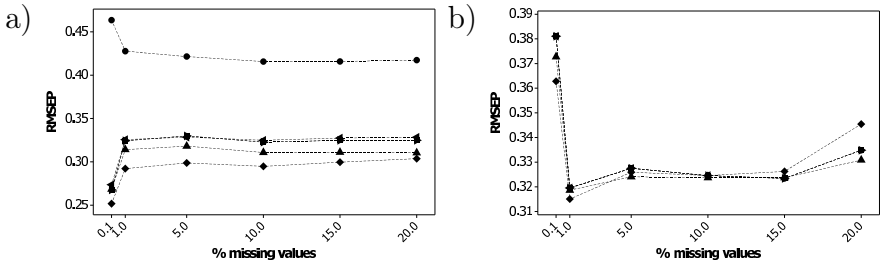

Figure 2: The minimum root mean square error of prediction (RMSEP) for different methods with different fractions of data missing for apo AI (a) and V583 (b). Average (●), FAR (■), KNNgene (♦), KNNobs (▲), PCR (►), and PLSR (◄).

there seems to be almost no change in RMSEP with increasing number of missing values, except when the fraction of missing values increases from 0.1% to 1%. The differences between Average and the other methods are constant over different fractions of missing values, except 0.1%.

From Figure 2a the non-parametric methods (KNN) seem to give best results, with KNN with genes as neighbours as the superior. The method FAR gave approximately the same results as PCR, and was also close to PLSR. All the regression methods gave better results than averaging over the gene. Regarding Figure 2b approximately equal results were obtained by all methods.

The number of components leading to the minimum RMSEP (Figure 2), are given in Table 1. For the regression methods the number of components

	apo AI					
Method	0.1%	1%	5%	10%	15%	20%
PCR	3	2	2	2	2	2
PLSR	1	1	1	1	1	1
FAR	3	2	2	2	2	2
KNNobs	4	4	4	4	5	4
KNNgene	8	11	18	21	21	29
	V583					
Method	0.1%	1%	5%	10%	15%	20%
PCR	4	4	4	4	4	4
PLSR	3	2	3	3	*	*
FAR	4	4	4	4	4	4
KNNobs	2	3	4	4	4	4
KNNgene	4	4	2600	1083	805	448

Table 1: Number of components for each method for each fraction of missing values. When 15% and 20% of the values in V583 are missing $\mathbf{R}^t\mathbf{SR}$ in (2) becomes singular for PLSR.

at the minimum RMSEP seems to be independent of the fraction of missing values, that means the fraction of missing values is of little importance for the decision of number of components. For the non-parametric methods (KNN) the number of components seems to change with increasing fraction of missing values. Note that minimum RMSEP is obtained by different number of components for the two data sets.

Figure 3 shows the behaviour of the methods at different numbers of components/neighbours when 1% of the data are missing. The results of

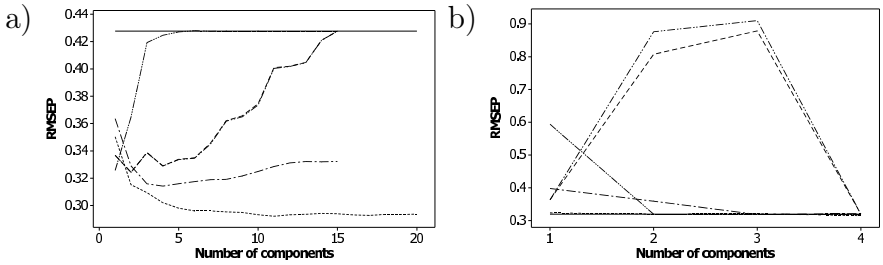

Figure 3: The RMSEP for different methods with 1% of the data missing for apo AI (a) and V583 (b). Average (——), FAR (— — —), KNNgene (- - -), KNNobs (— - —), PCR (— - - —), and PLSR (— - - - —).

the other fractions are not shown here, but these results are almost identical. For V583 we achieved the best results using gene average, while the opposite was the case for apo AI. In the latter case few components for the regression methods gave better results than a large number of components. The methods obtained their minimum with 1-2 components (Table 1). On the

contrary, using KNN the number of neighbours has almost no relevance as long as we use enough neighbours.

The results in this paper are based on one realization of the data set at each fraction of missing values. To study the effect of which values are missing, five different realizations of the data set of 1% missing values for apo AI were done (results not shown). The methods and the number of components were ranked in the same way in all the realizations, hence the choice of method seems independent of the data randomly chosen as missing. There were a slight difference in the RMSEP for the different realizations (the standard deviation of RMSEP was approximately 0.012). The standard deviation of RMSEP will decrease with increasing number of missing values (the standard deviation of RMSEP was approximately 0.006 for 20% missing values), since RMSEP will be based on a higher fraction of the data and different realizations may have more of the missing values in common.

To decide if the methods gave significantly different error of prediction, i.e. there exists an optimal method with an optimal number of components, we used the model in equation (9) on responses received from FAR, PCR, PLSR and KNN with observations as neighbours. This model was also used to decide if the optimal method and the optimal number of components were different for each gene. Table 2 presents the p-values for different tests based on this model. (Since square error is probably non-normal, an

Source	p-value (apo AI)	p-value (V583)
method	<0.0001	0.0089
component(method)	<0.0001	<0.0001
gene	<0.0001	<0.0001
gene*method	<0.0001	<0.0001
gene*component(method)	>0.9999	<0.0001

Table 2: Table of p-values corresponding to the tests based on model (9).

analysis based on absolute error was carried out. However the p-values were almost identical (results not shown).)

From Table 2 it is clear that both method and component within method were significant. There was also a significant effect of gene and of the interaction between gene and method, which means that the optimal method for one gene is not necessarily the optimal method for another gene. In apo AI we could not prove any effect of interaction between gene and number of components within each method, this is due to large variance inside the genes. In V583 we found this interaction to have significant effect.

5 Discussion

We have studied different methods for replacing missing values. In addition to previously studied methods; PCR, PLSR, KNN with genes as neighbours, and gene average, we focused on two methods; FAR and KNN with observations as neighbours. In addition we applied a method based on linear mixed models to compare the prediction methods.

This study considers situations where data are missing completely at random, i.e. the mechanism responsible for the missing values is not influenced by the values of the variables. Another situation, beyond the scope of this paper, is when data are missing at random, i.e. the missing values are independent of the gene expression, but dependent of some other variables, e.g. the corresponding background value.

The methods studied in this paper have been applied to two different data sets, one with relatively strong correlation structure between the genes and one with weak correlation structure between the genes. The strength of the correlation is reflected by the proportions between the eigenvalues of the covariance matrix. If there is strong correlation structure, one or several of the eigenvalues are relatively large. In the case of weak correlation structure, all eigenvalues are approximately equal. In the latter case the average is assumed to be the best predictor. For data with stronger correlation structure, the regression methods or the non-parametric methods should be applied, since they use the information among genes that lies in the correlation. The non-parametric methods (KNN) seem to give best results, with KNN with genes as neighbours as the superior. This can be explained by the assumption that there exist groups of coregulated genes, hence using genes from the same group as neighbours we achieve good prediction.

Intuitively we might assume that an increased error of prediction (defined as θ in (7)), and hence RMSEP, should follow from an increasing number of missing values, due to the fact that we have less information. If this is not the case, it indicates that the error of prediction differs for some genes. When 0.1% of the values were missing, approximately 1.5% (apo AI) and 0.5% (V583) of the genes had missing values, and the estimate of the prediction error was sensitive to which genes had missing values due to the effect of gene. When more genes had missing values, the estimate was more robust, and it stabilized. If the prediction ability among the methods for different fractions of missing values was non constant, it would be an indication of an interaction effect between gene and method.

In this paper we have studied the ability of different methods to predict missing values in gene expression matrices. However, the final interest is not the predicted values themselves, but rather how they influence on the further analysis. The optimal prediction method is the method that gives similar conclusions as we would obtain by analyzing the original data. Another criteria than RMSEP could have been better suited for this purpose. To improve the prediction for further analysis of the data,

the accuracy of the methods could have been investigated over the range of the expression values (Nguyen *et al.*, 2004). However, this is beyond the scope of this paper.

Our final goal is to predict the gene expression values originally missing in the gene expression matrix. Those genes are never used in the calibration of the model, since they were removed to receive a complete matrix. The significant effect of gene indicates that there exists no general error of prediction for all the genes, hence it is hard to estimate the error of prediction for the genes that originally have missing values. In situations with significant effect of the interaction between gene and number of components within method, there exists no number of components/neighbours optimal for prediction of every gene. When predicting the genes that originally missed values, we do not know what method is best for those genes. With no interaction, we assume that the optimal number of components in the calibration is also the best on the genes originally left out.

Our study shows that the optimal prediction method and number of components heavily depend on the gene expression matrix, hence there is no general advice for all situations. The first step should always be an investigation of the eigenvalues of the covariance matrix, for the purpose of achieving some knowledge of the correlation structure, and thereby the most suitable prediction methods. Weak correlation structure indicates that average is the optimal method, while stronger correlation structure indicates that regression methods and KNN are better choices. Stronger correlation structure requires a smaller number of components than weaker correlation structure. The best method for each situation can be found by first removing all genes with missing values, yielding a complete matrix. Further on, one removes from the complete matrix approximately the same fraction of data originally missing, and tries out different methods and number of components on the new matrix. The mixed model proposed in (9) can then be used to test which methods and number of components that are best for each gene, and if the differences are significant. A new realization of the data set should be used to estimate the corresponding error of prediction. Finally one finds the combination of methods and components that gives the best prediction and uses it on the original matrix.

References

[1] Almøy, T. (1994). Prediction methods with many possible explanatory variables. *PhD Thesis, Department of Mathematics, University of Oslo, Norway.*

[2] Buck, S.F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society Series B - Statistical Methodology*, **22**, 302-306.

[3] Bø, T.H., Dysvik, B., and Jonassen, I. (2004). LSImpute: accurate estimation of missing values in microarray data with least squares methods. *Nucleic Acids Research*, **32**, e34.

[4] Callow, M.J., Dudoit, S., Gong, E.L., Speed, T.P., and Rubin, E.M. (2000). Microarray expression profiling identifies genes with altered expression in HDL deficient mice. *Genome Research*, **10**, 2022-2029.

[5] Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O., and Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast. *Science*, **282**, 699-705.

[6] Dempster, A.P, Laird, N.M, and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with Discussion). *Journal of the Royal Statistical Society Series B - Statistical Methodology*, **39**, no. 1, 1-38.

[7] DeRisi, J., Penland, L., Brown, P.O., Bittner, M.L., *et al.* (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics*, **14**, 457-460.

[8] Efron, B. (1994). Missing data, imputation and the bootstrap. *Journal of the American Statistical Association*, **89**, 463-479.

[9] Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 14863-14868.

[10] Ernst, L.R. (1980). Variance of the estimated mean for several imputation procedures. *American Statistical Association, Proceedings of the Survey Research Methods Section*, 716-720.

[11] Ghosh, D. (2003). Penalized discriminant methods for the classification of tumors from gene expression data. *Biometrics*, **59**, 992-1000.

[12] Hacia, J.G. (1999). Resequencing and mutational analysis using oligonucleotide microarrays. *Nature Genetics*, **21**, 42-47.

[13] Hastie, T., Tibshirani, R., Sherlock, G., Eisen, M., Brown, P., and Botstein, D. (1999). Imputing missing data for gene expression arrays. *Technical report, Stanford Statistics Department.*

[14] Healy, M.J.R., and Westmacott, M. (1956). Missing values in experiments analyzed on automatic computers. *Applied Statistics - Journal of the Royal Statistical Society Series C*, **5**, 203-206.

[15] Helland, I.S., and Almøy, T. (1994). Comparison of prediction methods when only a few components are relevant. *Journal of the American Statistical Association*, **89**, 583-591.

[16] Höskuldsson, A. (1988). PLS regression methods. *Journal of Chemometrics*, **2**, 211-228.

[17] Indahl, U.G., and Næs, T. (1998). Evaluation of alternative spectral feature extraction methods of textural images for multivariate modeling. *Journal of Chemometrics*, **12**, 261-278.

[18] Lawley, D.N. and Maxwell, A.E. (1973). Regression and factor analysis. *Biometrica*, **60**, 331-338.

[19] Lindman, H.R. (1992). *Analysis of Variance in Experimental Design*. Springer-Verlag, New York.

[20] Little, R.J.A. and Rubin, D.B (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

[21] Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979). *Multivariate Analysis*. Academic Press, London.

[22] Martens, H. and Næs, T. (1989). *Multivariate Calibration*. John Wiley & Sons, New York.

[23] Massy, W.F. (1965). Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, **60**, 234-256.

[24] Nguyen, D.V., Arpat, A.B., Wang, N., and Carroll, R.J. (2002a). DNA microarray experiments: biological and technological aspects. *Biometrics*, **58**, 701-717.

[25] Nguyen, D.V., and Rocke, D.M. (2002b). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, **18**, 39-50.

[26] Nguyen, D.V., Wang, N., and Carroll, R.J. (2004). Evaluation of missing value estimation for microarray data. *Journal of Data Science*, **2**, 347-370.

[27] Næs, T., and Helland, I.S. (1993). Relevant components in regression. *Scandinavian Journal of Statistics*, **21**, 239-250.

[28] Næs, T., Irgens, C., and Martens, H. (1986). Comparisons of linear statistical methods for calibration of NIR instruments. *Applied Statistics - Journal of the Royal Statistical Society Series C*, **35**, 195-206.

[29] Ouyang, M., Welsh, W.J., and Georgopoulos, P. (2004). Gaussian mixture clustering and imputation of microarray data. *Bioinformatics*, **20**, 917-923.

[30] Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

[31] Rubin, D.B., and Schenker, N. (1991). Multiple imputation in health-care databases: an overview and some applications. *Statistics in Medicine*, **10**, 585-598.

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

[33] Wu, C.F.J. (1983). On the convergence of the EM algorithm. *Annals of Statistics*, **11**, 95-103.

[34] Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. *Empire Journal of Experimental Agriculture*, **1**, 129-142.