Computational Statistics and Data Analysis 53 (2009) 1650–1660

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Distribution modeling and simulation of gene expression data

Rudolph S. Parrish [a,*], Horace J. Spencer III [b], Ping Xu [a]

[a] Department of Bioinformatics and Biostatistics, School of Public Health and Information Sciences, University of Louisville, 555 S. Floyd St, Suite 4026, Louisville, KY 40292, USA

[b] Department of Biostatistics, College of Public Health, University of Arkansas for Medical Sciences, West Markham St., Little Rock, AR, USA

ARTICLE INFO

Article history:
Available online 29 March 2008

ABSTRACT

Data derived from gene expression microarrays often are used for purposes of classification and discovery. Many methods have been proposed for accomplishing these and related aims, however the statistical properties of such methods generally are not well established. To this end, it is desirable to develop realistic mathematical and statistical models that can be used in a simulation context so that the impacts of data analysis methods and testing approaches can be established. A method is developed in which variation among arrays can be characterized simultaneously for a large number of genes resulting in a multivariate model of gene expression. The method is based on selecting mathematical transformations of the underlying expression measures such that the transformed variables follow approximately a Gaussian distribution, and then estimating associated parameters, including correlations. The result is a multivariate normal distribution that serves to model transformed gene expression values within a subject population, while accounting for covariances among genes and/or probes. This model then is used to simulate microarray expression and probe intensity data by employing a modified Cholesky matrix factorization technique which addresses the singularity problem for the "small n, big p" situation. An example is given using prostate cancer data and, as an illustration, it is shown how data normalization can be investigated using this approach.

1. Introduction

In high-dimensional gene expression microarray systems, one of the important problems facing investigators concerns development of new statistical methods (or efficient application of existing methods) to achieve aims of classification or discovery. One might wish to classify an individual patient's disease so that the patient's therapy could be tailored for maximal efficacy, for example. Determining which genes play a role in the disease is a process of discovery involving microarrays as a screening tool. Another, less often discussed, objective is to monitor disease progression or recovery subsequent to diagnosis or treatment. Gene expression relates to protein synthesis and metabolic parameters in a systems biology context, so that characterization at the level of the gene is useful in focusing subsequent efforts, such as investigating gene regulatory networks and pathways. Ultimately, one needs to reduce the vast information contained in microarrays, and perhaps other relevant clinical parameters, to a relatively simple form which permits its practical use. Statistically, this includes characterizing the data as succinctly as possible without losing information content. Mathematical and statistical modeling methods can be applied advantageously to help address these aims. If microarray data can be modeled accurately, statistical methods can be examined with respect to their relative usefulness in applications. This includes various methods relating to significance testing for differences between groups and identification of differentially expressed genes. While

[*] Corresponding author. Tel.: +1 502 852 2797; fax: +1 502 852 3294.
E-mail address: rsparr01@gwise.louisville.edu (R.S. Parrish).

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.03.023

there are several proposed methods for doing this, there is very little information about the performance characteristics of such methods when applied to microarray data.

In experiments involving different subjects (or, equivalently, multiple independent samples), there is natural subject-to-subject variation that occurs for expression levels of all genes (usually termed "biological variation"), and there is natural variation that occurs when multiple arrays are utilized within the same subject which can be due to sampling error or other factors. Variation among subjects is the basis for the experimental error term used in significance testing. That is, the subject generally is considered as being the experimental unit. Other situations could involve multiple samples derived longitudinally from the same subject (e.g., time-course experiments). Systematic or "technical" variation also often exists in relation to the experimental design or other aspects of conducting the experiment.

Singhal et al. (2003) developed a microarray data simulator in which systematic and random technical variability and biological variability were added onto a model of underlying gene expression. Nykter et al. (2006) developed a modular model involving sources of variation that included slide manufacturing, biological noise, hybridization, and scanning error; this model requires specification of a large number of parameters. Albers et al. (2006) provided a web page for simulation of two-channel cDNA arrays. While addressing underlying sources of variation, none of these approaches incorporated correlations among genes, and all relied on multiple assumptions. In the course of investigating properties of various methods, many other authors (e.g., Molinaro et al. (2005)) have used simulation approaches typically based on the multivariate normal distribution with fixed, simplified covariance structures. In reality, the covariance matrix does not follow a simple structure, nor do the marginal distributions follow a normal or even lognormal model. Thus, something more realistic is needed.

In this paper, a method is proposed in which variation among arrays (usually one array per biological unit) can be characterized simultaneously for a large number of genes resulting in a multivariate model of gene expression, which incorporates both biological and technical variation. The emphasis is on characterizing the true nature of the array data without pre-specification or presumption of an underlying, unknown 'ground truth', as other methods typically require. In some contexts, the 'ground truth' can be thought of as the parent multivariate distribution model that gives rise to the data. It is possible to incorporate treatment effects and other sources of variation if the application requires it.

The proposed method is based on selecting mathematical transformations of the underlying expression measures such that the transformed variables follow approximately a Gaussian (i.e., normal) distribution, and then estimating associated parameters, including correlations (or covariances). The result of this procedure is a multivariate normal distribution that serves to model the transformed gene expression values among subjects and takes into account correlations among genes. The same approach can be used to characterize underlying probe-level intensity data, incorporating correlations also among probes within genes. This approach makes it possible to simulate gene expression and probe intensity data enabling an investigator to produce arbitrarily large sample sizes so that properties of different statistical methods can be evaluated and compared. This includes normalization, classification (Xu et al., 2009), and differential-expression testing algorithms. Another potential use is in development of numerical indices that relate to disease progression and recovery when disease status is assessed longitudinally. With this approach, epidemiologically derived gene expression data can be summarized for populations of interest to be used as reference groups. In addition to use of gene expression data, this approach also makes it possible to include clinically derived information for continuously measured variables and their correlations with gene expressions. In any given situation, subsets of genes can be modeled using the corresponding joint distribution derived from the complete model fit, thus enabling detailed analyses of genes of interest, as for example those in some defined pathway.

In the following, we first describe the methodology for fitting the multivariate model and subsequently generating random array data. We then evaluate the method by comparing various aspects of simulated data to the training data, particularly as regards the goodness of fit to the empirical distributions and the capability to reproduce correlations among genes and probes. The results are based on application to data derived from a study of prostate cancer patients. An example of applying this methodology to the problem of data normalization is presented for illustration.

2. Methods

2.1. *Affymetrix GeneChip® technology*

Technical aspects of oligonucleotide arrays have been discussed in several publications (Affymetrix, 2002; Nguyen et al., 2002). In brief, GeneChip® arrays provide information on gene expression by measuring multiple probe intensities, each consisting of 25-mer base pairs or more, depending on the version of the array design, which then are utilized in one of many available algorithms to produce composite measures of gene expression. The Human U95Av2 array contains a grid of 640 × 640 probes grouped in 12,525 sets, with each set consisting of 11–20 probe pairs (median 16), and with each set corresponding to a single gene transcript. Some microarrays map multiple probe sets to individual genes. Commonly employed expression algorithms, which include Microarray Suite 5.0 (MAS 5.0) (Affymetrix, 2002), Robust Multiarray Average (RMA) (Irizarry et al., 2003), and dChip (Li and Wong, 2001), mathematically process probe-level data to produce gene-specific expression values.

2.2. Multivariate normal distribution model

In order to develop a multivariate distribution model of gene expression, expression data are examined for each gene individually using an automated algorithm to select a transformation that produces approximately a Gaussian (i.e., normal) probability distribution. For purposes of this discussion, arrays are assumed to be independently associated with individual subjects, although generalization to multiple arrays within subjects is straightforward. For each gene, a transformation is determined on the basis of data from all subjects in a particular category, such as healthy subjects or patients with disease. After making all appropriate transformations for the entire set of genes under study, the means, variances, and covariances (between pairs of genes) of transformed variables are estimated. Thus, a multivariate Gaussian distribution model is constructed having as density function the general form: $f(\mathbf{z};\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\{-0.5(\mathbf{z}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})\}$, $-\infty < z_i < \infty$, where $\mathbf{z}$ is the vector of transformed variables, $\boldsymbol{\mu}$ is the vector of means of the transformed variables, and $\boldsymbol{\Sigma}$ is the matrix of variances and covariances of the transformed variables. In essence, this process involves selecting appropriate transformations that result in marginal distributions of the transformed variables which are approximately Gaussian.

An important property of this model is that any subset of the transformed variables also has a multivariate normal distribution with corresponding parameters taken directly from $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. In addition, there are closed-form relations for determining the corresponding variance–covariance matrix in the case where conditional distributions are of interest (Morrison, 1976; Johnson and Kotz, 1972). This means that subsets of genes are automatically modeled as soon as the overall model is fully specified, so that genes which may be related to each other as a result of a common gene regulatory network or pathway can be considered independently of other genes.

2.3. Transformations to normality

Based on our extensive empirical investigations, no single transformation family can provide the scope of transformations necessary to obtain a high percentage of good approximations to the data, as measured by a goodness-of-fit criterion. Often a simple lognormal transformation is employed as a normalizing transformation, but it fails to provide adequate fits in a substantial proportion of cases, sometimes for as much as 70% of the data. In order to find a best fitting distribution, two types of normalizing transformation families were considered: (1) Box–Cox power transformation (Box and Cox, 1964) and (2) Johnson transformation system (Johnson and Kotz, 1970). Both of these are nonlinear methods and may require iterative algorithms for implementation. The Johnson system includes the lognormal distribution. The Box–Cox transformation was recently considered by Giles and Kipling (2003) for use with microarray data. Some modifications to this method have been proposed (see Sakia (1992), for a review and bibliography, and also Kirisci et al. (2005)), however the original transformation was used here. In our work, an empirical approach to fitting was adopted in which each possible transformation was applied to each gene's data (across subjects) and the Kolmogorov–Smirnov (KS) goodness-of-fit (GOF) statistic was used to identify the best fitting transformation in each case. These transformations should be considered as mechanisms to achieve approximate normality. They are essentially four-moment transformations, thus issues of overfitting are not of concern. For example, with the Johnson system, skewness and kurtosis can be used to define which member of the family is selected for modeling a marginal distribution. In a diagram presented in Johnson and Kotz (1970), the lognormal distribution corresponds to a curved line in the skewness–kurtosis plane, which implies that the family of distributions is much richer than the lognormal distribution allows, and hence distribution modeling is more likely to succeed. After transformation, all that is needed for the multivariate model are the means, variances, and covariances, as well as the limits of variation for some distributions. With the Box–Cox family, a single parametric transformation is employed but requires an iterative approach. In any given case, a distribution from either family could be better than the other.

(1) *Box–Cox transformation.*

The traditional Box–Cox family of transformations is defined by $y = (x^\theta - 1)/\theta$, if $\theta \neq 0$, and $y = \log(x)$, if $\theta = 0$, where θ is a parameter selected to achieve normality and where x is the original (i.e., untransformed) expression value of a given gene transcript, as described above. Estimation of θ is based on maximization of the correlation coefficient between the x- and y-values of a standard normal probability plot. Computationally, an optimal value of θ can be found by an iterative search process.

(2) *Johnson system of distributions.*

The Johnson system of distributions is defined on the basis of three transformations; these are denoted by: S_L (lognormal), S_B (bounded), and S_U (unbounded). In addition, a null transformation is denoted here by S_N, where the subscript denotes "normal". This system uniquely associates a distribution with each possible pair of values of skewness ($\sqrt{\beta_1}$) and kurtosis (β_2), the standardized third and fourth moments ($\sqrt{\beta_1} = \mu_3/\mu_2^{3/2}$ and $\beta_2 = \mu_4/\mu_2^2$, where μ_r represents the rth central moment). That is, each possible pair of skewness and kurtosis values is associated with a unique distribution in the Johnson system. Possible (β_1, β_2) values are restricted to the region $\beta_2 \geq \beta_1 + 1$ (Johnson, 1987). The normal distribution has fixed skewness and kurtosis coefficients equal to 0 and 3, respectively, and this is represented as a single point in the skewness–kurtosis plane. The lognormal distribution has skewness and kurtosis coefficients that are related according to the parametric equations $\beta_1 = (\omega - 1)(\omega + 2)^2$ and $\beta_2 = \omega^4 + 2\omega^3 + 3\omega^2 - 3$ where $\omega = \exp(\eta^2)$ with η^2 being the variance of the log-transformed variable. Thus, in the (β_1, β_2) plane, these equations define a curved line that divides the possible pairs of values. On one side of this line, corresponding to kurtosis lower than for lognormal (for fixed skewness), are points corresponding to S_B distributions; while the other side corresponds to S_U distributions and has kurtosis higher than for the lognormal. As the Johnson system of distributions includes combinations of skewness and kurtosis that are far different from the lognormal, it provides a much wider variety of normalizing transformations than are available with the log transformation alone. A good approximation is more likely to be found using the Johnson system than if the lognormal distribution were used exclusively.

Letting x represent a gene expression variable, a standardizing transformation for location and scale may be applied in the form: $(x-\xi)/\lambda$ where ξ and λ are constants. For the S_L distribution, the value of λ may be taken to be 1 without loss of generality; and thus ξ is the lower terminus of the three-parameter lognormal distribution. For the S_B distribution, λ will be the range of variation, $\lambda = B - A$ where $A < x < B$, and $\xi = A$ is the lower limit of variation. The Johnson normalizing transformations, as given by Rose and Smith (2002), are defined as:

$$S_L: \; y = \log(x-\xi), \quad \xi < x < \infty, \tag{1}$$

$$S_B: \; y = \log[(x-\xi)/(\xi+\lambda-x)], \quad \xi < x < (\xi+\lambda), \tag{2}$$

$$\begin{aligned} S_U: \; y &= \sinh^{-1}[(x-\xi)/\lambda] \\ &= \log[(x-\xi)/\lambda + \{1+[(x-\xi)/\lambda]^2\}^{1/2}], \quad -\infty < x < \infty, \end{aligned} \tag{3}$$

and

$$S_N: y = (x-\xi)/\lambda, \quad -\infty < x < \infty. \tag{4}$$

Further define $z = \gamma + \delta y$ where the parameters γ and δ control the shape of the distribution. In each case, parameter values are sought so that Z has a standard normal distribution. Thus, this system involves four parameters: ξ, λ, γ, and δ. For estimation purposes, let $\mu = -\gamma/\delta$ and $\sigma = 1/\delta$. These correspond to the mean and standard deviation of Y for one of the four transformations, as the case may be. With these definitions, $y = \mu + \sigma z$.

2.4. Estimation of means, variances, and covariances

Methods using iterative techniques are available for estimating the parameters of the distributions based on the Box–Cox and some members of the Johnson transformations (Rose and Smith, 2002; Johnson, 1949, 1965; Johnson and Kitchen, 1971; Zhou and McTague, 1996). For the Johnson system, methods based on percentiles have been described by Wheeler (1980), Slifker and Shapiro (1980), Bowman and Shenton (1989), and Shayib (1989). The method of moments also may be used in some cases. Although each method proposed for transforming to normality has advantages, none seems to be clearly superior, especially considering the variability seen among the large number of genes involved. Thus, an empirical approach is taken here for obtaining a best fitting normal approximation in which the KS GOF statistic is used as the fitting criterion. Each of the approaches described below is employed in turn and the corresponding numerical value of the KS statistic is obtained. The final selection of a transformation and the associated parameter estimates correspond simply to the minimum-valued KS statistic that is observed. This is done for each gene so that the end result is a table containing designations of the type of transformations and the associated parameter estimates.

(a) *Box–Cox transformation:* An initial value for the parameter θ in the Box–Cox formulation is obtained by a systematic search over the interval [−3,3] in equal steps (e.g., 15 values). For each such value, the Box–Cox transformation is made, moment estimates are used for the normal distribution parameters, and the KS GOF statistic D is evaluated. The two smallest consecutive values of D determine an interval that contains the θ value which produces the minimum D. A binary search algorithm is then employed to converge on the optimal solution for θ, in which each step requires that the transformation be made, distribution parameters be estimated, and the GOF criterion D be computed. Throughout this procedure, if the absolute value of θ is less than a specified small tolerance value, such as 0.05, the log transformation is applied rather than the power transformation. At each step, after the Box–Cox transformation is done, the sample mean (m_1) and standard deviation (m_2) for the transformed values are computed, which are used to compute the parameter estimates: $\gamma^* = -m_1/m_2$ and $\delta^* = 1/m_2$.

(b) *Johnson system:* Selection of an appropriate Johnson distribution may be based on the third and fourth standardized moments (i.e., skewness and kurtosis), followed by an iterative approach for estimating distribution parameters for the selected type. The large number of genes and potential numerical problems make this approach difficult to implement reliably, especially since higher-order moments are more difficult to estimate with small sample sizes. Although the choice of distribution type can be based on the skewness and kurtosis values, anomalies in the data that translate to variability in moment estimates sometimes make it difficult to pick reliably the best fitting approximation using this method. Owing to these factors, the selection and fitting process for Johnson distributions employs both the method described by Slifker and Shapiro and an empirical-based *ad hoc* approach as described below.

Slifker–Shapiro fitting method: The method of Slifker and Shapiro (1980) is based on the use of four equally spaced points from the transformed standard normal distribution, denoted by $-3z$, $-z$, z, and $3z$. When $z = 0.524$, for example, this corresponds to the 70th percentile and $3z = 1.572$ corresponds to the 94.2th percentile. Different values of z may be more desirable for use with small data sets, such that $3z$ is not at the extreme quantiles of the empirical data. The empirical percentiles are derived from the raw data by calculating the cumulative standard normal probabilities for these four points (denoted by P_{-3z}, P_{-z}, P_z, and P_{3z}) and then inferring the corresponding raw values using the ith order statistic of the raw data where

$i = NP + 0.5$, with N being the sample size. Interpolation is used when i is not an integer. These empirical percentiles are denoted by x_{-3z}, x_{-z}, x_z, and x_{3z}. Letting $m = x_{3z} - x_{-z}$, $n = x_{-z} - x_{-3z}$, and $p = x_z - x_{-z}$, the following ratios determine the type of Johnson distribution:

$S_U : mn/p^2 > 1, S_B : mn/p^2 < 1$, and $S_L : mn/p^2 = 1$. As the S_L ratio is unlikely ever to be exactly equal to unity, a tolerance interval around 1 is used to permit selection of the lognormal. (For this analysis, a tolerance of 0.05 was used.) Then, the distribution parameters are estimated using the closed-form formulas given by Slifker and Shapiro.

Empirical fitting method: In the empirical fitting method, a GOF-based procedure is applied for each gene individually and for each of the Johnson distributions separately. After all distribution types have been optimally fitted in this fashion, the one which has the smallest KS GOF statistic is selected as the best fitting Johnson distribution. The steps are as follows:

1. Select one of the Johnson distribution transformations.
2. Determine initial values A and B to represent the range of possible values for X. The absolute limits for A are 0 to $x_{\min}$, and for B are $x_{\max}$ to $2^{16} - 1$ (or 16 if a log-base-two transformation is made to the raw intensity values). The following initial values are suggested: $A = X_{\min} - tol^*(x_{\max} - x_{\min})$ and $B = X_{\max} + tol^*(x_{\max} - x_{\min})$ for some pre-specified tolerance value, such as 0.05.
3. Apply the following procedure for the current values of A and B:
   (a) Set $\xi = A$ and $\lambda = B - A$, and compute y-values as given above for the selected transformation.
   (b) Calculate the sample mean (m_1) and standard deviation (m_2) of these transformed values for each gene, to be used as moment estimators of the normal distribution parameters.
   (c) Compute the one-sample KS GOF test statistic, D, used in assessing normality.
4. Vary the tolerance value over a defined range, such as 0.01 to 0.09, and modify the values of A and B until an optimum value for the goodness-of-fit statistic is achieved for each distribution type. The effect of this is to determine values for A and B corresponding to each distribution separately such that the KS statistic is minimized.
5. Select the transformation corresponding to the smallest value of D for each gene.
6. Compute the sample mean (m_1) and standard deviation (m_2) for the selected transformation values (y). These optionally can be based on robust procedures such as trimmed moments.
7. The parameter estimates (denoted by *) then are given by: $\xi^* = A$, $\lambda^* = B - A$, $\gamma^* = -m_1/m_2$, $\delta^* = 1/m_2$.

For simplicity of fitting, ξ is taken as the lower limit of variation of x and $\xi + \lambda$ is taken as the upper limit, although these may not be optimal choices due to the transformations involved. While this procedure is not statistically optimal, it performs adequately for the purpose of selecting and fitting a distribution as an approximation to the underlying true distribution, and it can be performed efficiently in the context of a large number of genes. A more refined approach involving estimation of ξ and λ can be implemented but, then, four sample moments are required for estimation rather than two and a nonlinear optimization algorithm must be employed in which numerical convergence is not guaranteed.

2.5. Simulation of gene expression data

Random deviates from a multivariate normal distribution can be generated easily using a triangular factorization of the variance–covariance matrix, provided such a factorization exists. Particularly, if the covariance matrix $\mathbf{\Sigma}$ can be factored as $\mathbf{\Sigma} = \mathbf{TT}'$ where $\mathbf{T}$ is a lower triangular matrix and if $\mathbf{u}$ represents a vector of standard normal independent random deviates, then $\mathbf{y} = \boldsymbol{\mu} + \mathbf{Tu}$ defines a vector of random deviates from the desired multivariate normal distribution, having the desired covariance structure. In this representation, $\boldsymbol{\mu}$ is the mean vector for the transformed variables ($\mathbf{y}$) and $\mathbf{\Sigma}$ is the covariance matrix for the transformed variables.

The Cholesky decomposition algorithm (Kennedy and Gentle, 1980) may be used to obtain $\mathbf{T}$ when $\mathbf{\Sigma}$ is positive definite. With microarray data, however, $\mathbf{\Sigma}$ usually will be only positive semi-definite, owing to the large number of genes under consideration relative to the number of arrays (experimental units). In this case, a modified Cholesky algorithm may be employed to obtain $\mathbf{T}$ (Schnabel and Eskow, 1990, 1991, 1999) which, in effect, modifies $\mathbf{\Sigma}$ so that it is positive definite. The modifications are generally minimal.

After generating standard normal deviates using the matrix transformation above, conversion to original scaling is accomplished using inverse transformations for the marginal distributions, which produces random deviates on the original scale. For the Johnson distributions, the inverse transformations are derived algebraically from the transformations given above (Rose and Smith, 2002):

$$S_L : \ x = \xi + \mathrm{e}^y, \tag{5}$$

$$S_B : \ x = [\xi + (\xi + \lambda)\mathrm{e}^y]/(1 + \mathrm{e}^y), \tag{6}$$

$$S_U : \ x = \xi + \lambda \sinh(y) = \xi + \lambda[\mathrm{e}^y - \mathrm{e}^{-y}]/2, \tag{7}$$

and

$$S_N : \ x = \xi + \lambda y, \tag{8}$$

where $y = (z - \gamma)/\delta$. For the Box–Cox case, the inverse transformations are: $x = (1 + \theta y)^{1/\theta}$, if $\theta > 0$, and $x = \mathrm{e}^y$, if $\theta = 0$, where $y = (z - \gamma)/\delta$.

2.6. Modeling probe-level data

Notwithstanding computational limitations, probe-level data can be modeled using the same approach as described above for expression data. However, the dimension of the numerical problem is increased by a factor of 11–16 which may limit the number of genes that can be so represented. It is possible to reduce the requirement for extremely high dimensions by utilizing a nested modeling approach (not discussed here).

2.7. Simulation of differential expression

A nominal experiment may be simulated using two treatments, N subjects per treatment, and one array per subject, replicated M times. Treatment effects may be introduced based on an assumed p-value distribution, perhaps a mixture of uniform and beta distributions described by Allison et al. (2002). In order to simulate differentially expressed genes, a p value is generated randomly from the mixture distribution for each gene, and the corresponding t quantile is computed. Under the assumption that a t test would be conducted to test for differential expression, a constant c can be determined and then added to each of the gene expression or probe intensity values in one of the treatment groups, so that the same empirical t value would be obtained had the modified data been used in the calculation. This assumes that the mean or median of the intensity values in each probe set is used as the summary measure of gene expression when simulating probe-level data. This process may be applied before or after background correction and/or normalization of the parent data.

2.8. Computer resources

To obtain a benchmark on computing (i.e., CPU) time, 500 arrays were simulated using 4200 genes on a dual-processor 2 GHz Windows XP system with 4 GB RAM. Approximately 49 min of CPU time were required for completion, with about 25% of the time being spent on distribution fitting and about 75% on factorization of the covariance matrix and generation of multivariate normal data. Using a single node on a 64-bit Linux cluster parallel system with dual processors and 4 GB of memory per node, arrays with 12,625 genes have been successfully simulated.

3. Results

In this section, we describe a real data set and model it using the foregoing proposed method. After generating simulated data from the fitted model, we compare various aspects of the simulated data to the real data in order to show that the simulated data adequately represent the important features in the parent data set. These include consideration of the following: (1) goodness-of-fit statistics which should show a high percentage of very good fits to the marginal distributions; (2) quantiles (i.e., percentiles) of the real and simulated distributions which should match closely, as shown by quantile–quantile plots; (3) distributions of pair-wise correlations among genes and probes of the real and simulated data which should match closely; and (4) multidimensional scaling components of the real and simulated data which should be similar. Criterion (1) will ensure that the fits to the data for each gene or probe are adequate overall with a high proportion of closely fitting approximations. This can be seen in a cumulative distribution plot of the goodness-of-fit statistics. Criterion (2) will establish that the percentage points, taken over genes and or probes, compare favorably, showing consistency over all genes and/or probes, especially for the tail areas of the distributions. Criterion (3) will ensure that gene- and/or probe-specific correlations are closely approximated, indicating that the multivariate model for covariance is reasonable. Criterion (4) will show that the underlying structure of the data is preserved. Establishing all of these criteria will provide strong evidence that the simulated data have the same characteristics as the real data.

3.1. Data from a study in prostate cancer

Singh et al. (2002) described a study in which data were collected on prostatectomy patients with specimens being collected both from tumors and adjacent non-tumor prostate tissue (hereinafter called “normal” samples). There were 50 normal and 52 tumor samples, with 47 total paired samples. Expression profiles were derived using oligonucleotide microarrays (H95Av2) containing 12,525 genes. Data files were acquired from the authors’ web site. Raw probe-level data were accessed in the form of “CEL” files produced by Affymetrix analysis software. Those data were processed using *R* code from the Bioconductor software package module AFFY (Bioconductor, 2003) in order to produce expression values utilizing methods analogous to those employed by Affymetrix MAS 5.0 software. The reason for this approach was that Affymetrix MAS 5.0 output values were available only to one decimal place, which proved insufficient for the purpose of fitting distributions. Model adequacy was assessed and effects of normalization were considered in order to demonstrate utility of the methodology.

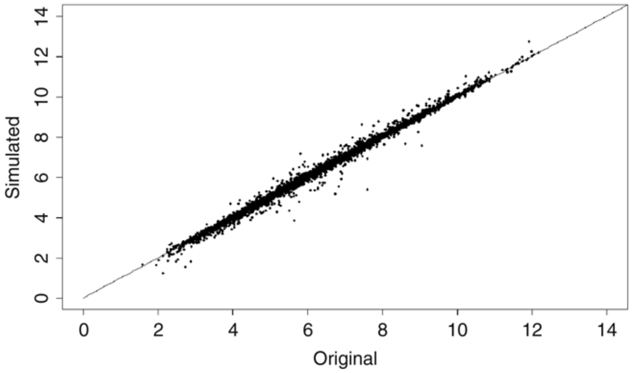

Fig. 1. Scatter plot of quantiles (0.01, 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95, 0.99) for original and simulated data based on MAS 5.0 expression values for 1000 genes.

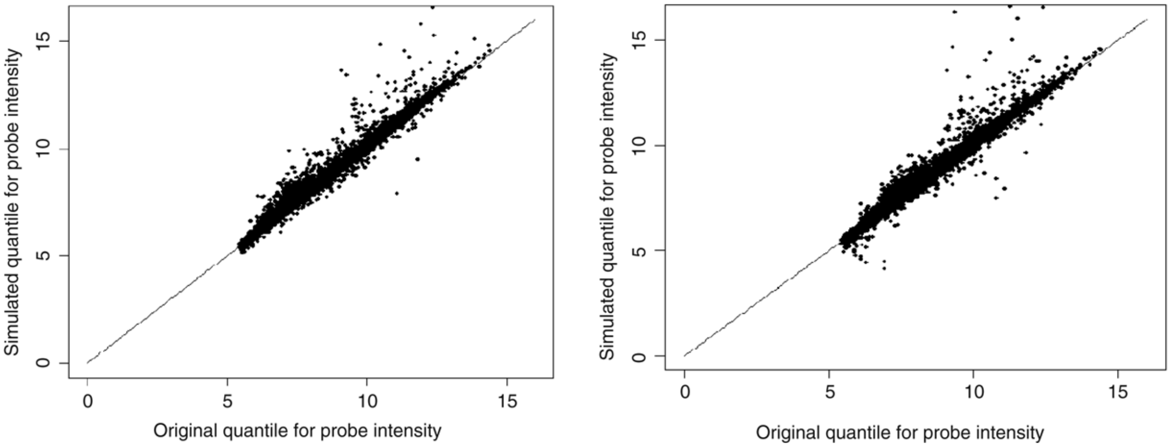

Fig. 2. Scatter plot of quantiles for original and simulated data for PM probes (left panel) and PM & MM probes combined (right panel).

3.2. Assessing model adequacy

The modeling and simulation procedures described above were applied to the data from normal tissue samples for three cases: (1) MAS 5.0 gene expression values, (2) individual PM probe intensities, and (3) individual PM and mismatch (MM) probe values. For analysis (1), 1000 genes were selected randomly. For probe-based analyses, case (2) used 200 genes with 16 PM probes each (i.e., 3200 probe values), and case (3) used 100 genes having 16 PM and 16 MM probes (i.e., 1600 PM and 1600 MM probe values). The success of the proposed modeling approach was examined for each of these cases by considering goodness-of-fit (GOF) statistics, distribution quantiles, and pair-wise correlations. Simulated data involved 500 arrays.

For case (1), the cumulative distribution functions (CDF) of KS GOF statistics show that 98% of the gene distributions were fitted with a KS GOF criterion of less than 0.125, which is acceptable when using the Gaussian distribution to approximate an underlying distribution. Results for other cases were similar. For expression data, approximately 15%, 32%, 1%, 1%, and 52% of the cases were fitted with S_U, S_B, lognormal, normal, and Box–Cox distributions, respectively. For probe intensity data, these were 13%, 32%, 20%, $<1\%$, and 36%. In all cases, a $\log_2$ transformation was applied before fitting so that data were scaled between 0 and 16.

There was close agreement between observed and simulated quantiles for both gene expressions (Fig. 1) and probe intensities (Fig. 2). Higher variation at the extreme quantile levels is expected, especially with smaller sample sizes. Fig. 3 shows histograms of the pair-wise correlations among genes for MAS 5.0 values for the observed data and the simulated data, indicating good agreement overall with some attenuation in the larger absolute values.

A plot of original versus simulated pair-wise correlations in the original scale following inverse transformation of the simulated data is shown in Fig. 4. There is excellent agreement among correlations. Histograms of correlations between original probe intensities and simulated values for PM and MM probes are given in Fig. 5. The bimodal nature of these plots is due to the mix of higher correlations among probes within the same genes and the correlations among genes.

As an additional graphical comparison, multidimensional scaling (MDS) was used to derive and plot the first two components (Hastie et al., 2001). For this analysis, quantile normalization and RMA summarization were employed on 50

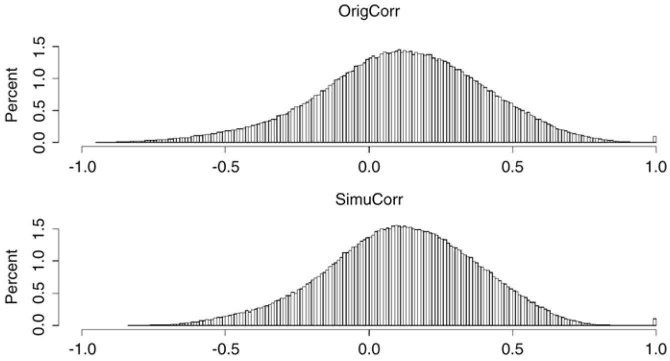

Fig. 3. Histograms showing distributions of pair-wise correlations among gene expression values for the observed data (top) and simulated data (bottom).

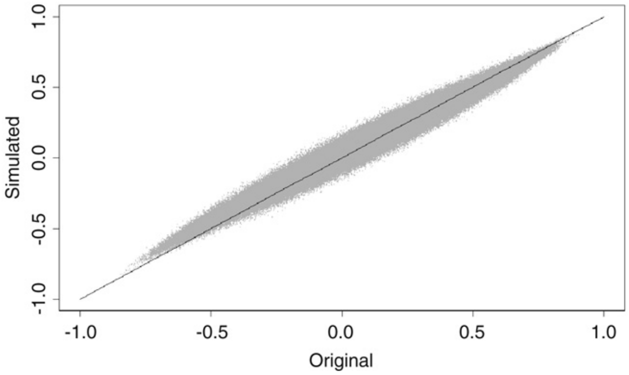

Fig. 4. Scatter plot of correlations computed from original data and from simulated data.

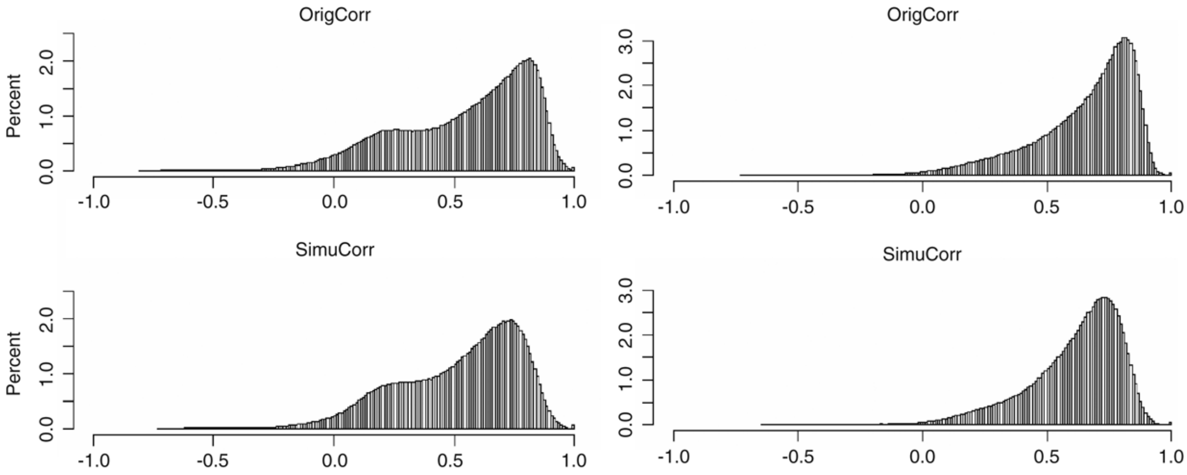

Fig. 5. Distributions of pair-wise correlations among PM probe values (left side) and MM probes (right side), showing results based on the original data (upper) and simulated data (lower).

normal tissue and 52 tumor tissue samples. The top 2000 differentially expressed genes were identified via empirical Bayes t-statistics for use in the MDS procedure. 500 normal and 500 tumor arrays were simulated. A plot is shown in Fig. 6, where the left panel is based on observed expression data and the right panel is based on simulated data. This indicates that the simulated data represent the primary structural components adequately.

3.3. Example: Using simulated data to examine the effect of data normalization

In order to illustrate how the simulation approach can be used to investigate properties of methods developed for analyzing microarray data, quantile normalization (Irizarry et al., 2003) was considered. In perhaps all cases, available

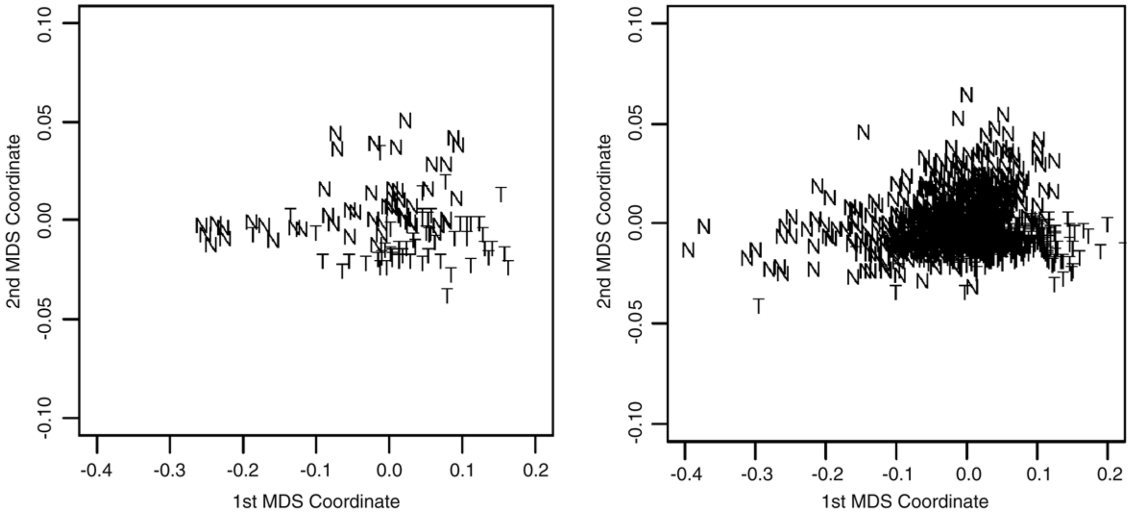

Fig. 6. Plot of MDS1 versus MDS2 for observed data from 102 arrays (left panel) and for 1000 simulated arrays (right panel). Symbols represent normal (N) and tumor (T) samples.

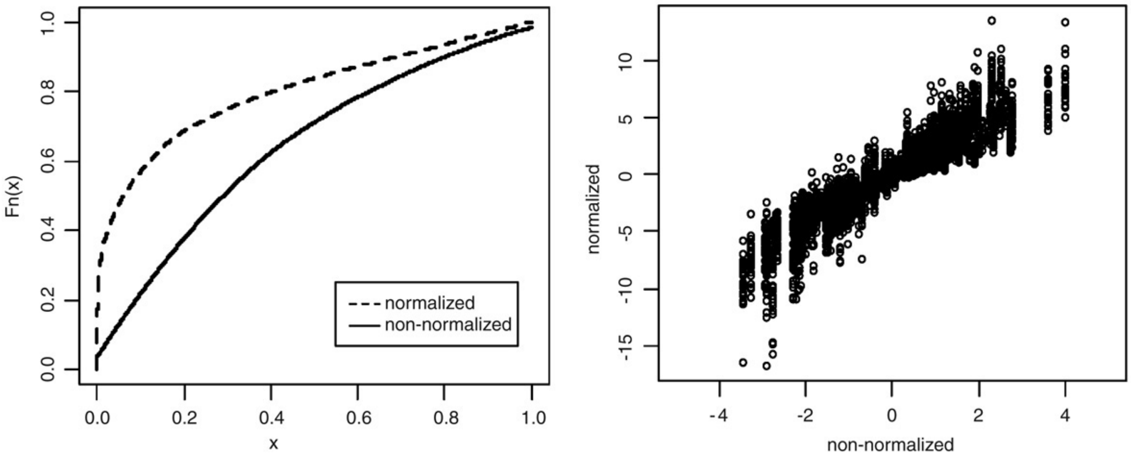

Fig. 7. Empirical CDF of p-values (left panel) and t statistics (right panel) for non-normalized and normalized data.

observed data are not sufficiently large in terms of number of subjects to allow one to establish statistical properties of a given normalization procedure. However, simulated data can be generated so that such features can be investigated. Based on using the prostate cancer data as a training set, probe-level data for 500 arrays and 200 genes (i.e., total of 3200 probes with 16 probes per gene) were simulated as described above. Interest was on examining the impact on p-values derived from t tests for treatment effect. A nominal experiment was simulated using two treatments, 10 subjects per treatment, and one array per subject, replicated 25 times. Treatment effects were introduced based on the p-value distribution

$$0.55 * \text{Uniform}(0, 1) + 0.45 * \text{Beta}(a = 0.775, b = 3.862) \tag{9}$$

which provided approximately 15% of values below 0.05. The constants used to introduce effects were computed according to the expression

$$c = t_{18,p}{}^{*}\sigma - (\mu_1 - \mu_2), \tag{10}$$

where $\mu_1$and μ_2 are the treatment means across subjects of the medians of probe intensities, σ represents the standard deviation of $\mu_1 - \mu_2$, and $t_{18,p}$ is the $(1 - p/2)$-level quantile of the t distribution with 18 degrees of freedom. Fig. 7 shows plots of p-values and t statistics using the generated data before and after quantile normalization. This indicates that a much higher proportion of large t-values (with small p-values) are calculated from the normalized data. For non-normalized data, only 2% of p-values were below 0.01, whereas, for normalized data, approximately 12%, 24%, and 44% were below 0.0001, 0.001, and 0.01, respectively.

4. Discussion

The proposed multivariate model for gene expression data is based on an empirical set of transformations to Gaussian form, with the intent of achieving a reasonable approximation to the true underlying distribution in every dimension. Although not discussed in this paper, our preliminary investigations revealed that a significantly higher percentage of acceptable fits overall (98%, based on a 0.125 KS GOF criterion) could be obtained by using more than one family of transformations, particularly the Box–Cox and Johnson systems. Our work indicated that, for typical data sets, a lognormal transformation alone might produce only about 30%–40% acceptable fits, whereas the Box–Cox approach alone or the Johnson family alone might achieve approximately 90%. The results for a cancer data set, and many others not discussed, show consistently that this procedure can adequately model the marginal distributions as well as the correlations among either gene expressions or probe intensities, including MM probes. By incorporating realistic variance–covariance information among genes or probes, this simulation approach can serve as a mechanism to investigate properties of various proposed methods for analysis of microarray data. This typically would involve numerically adding constants or perturbations to the genes or probes to reflect design effects. Particularly, normalization, summarization, and classification methods can be evaluated objectively. Since every subset of variables from the multivariate normal distribution also has a multivariate normal distribution, this technique may be useful as a tool to investigate gene networks and pathways. This model also could incorporate continuous clinical or other ancillary variables. This approach should be useful in situations where it is desired to have large data sets for establishing properties of microarray analysis methods without explicitly specifying all underlying relationships among genes and particular error structures. It can also be useful for developing a representation of large-scale multidimensional microarray data using relatively few parameters.

This approach has some limitations. Clearly, the true relationships among genes is not specified or otherwise assumed to be known *a priori*. Whatever those relationships are, they are assumed to be captured intrinsically through the multivariate model. Accepted methods for selecting approximating distributions and estimating the associated parameters are employed. A high level of flexibility is attained by not forcing a single distribution type to be used for all genes. If the distribution parameters can be considered as the 'truth', then one does in fact have some measure of the 'ground truth' even though it is not completely specified from a gene network standpoint. Another limitation is the size of data sets upon which the model is based. Typical microarray experiments do not involve large numbers of samples, thus the estimated multivariate distribution will be predictably impacted. Of course, larger data sets are more likely to result in good models of the underlying process, and one might argue that as the technology becomes less costly the size of data sets also will increase according to the needs of the application. The method may be most appropriate for studying problems where it is important to capture the underlying relationships intrinsically or to simply characterize populations of interest in a relatively succinct manner. In addressing how robust the modeling process is for a given purpose, standard methods (e.g., cross validation) could be employed to investigate performance features of the method. Although the available computer hardware may impose limitations on the number of genes that can be represented, we have found for the applications of interest that this number is sufficient to carry on methodological investigations. In other words, for such purposes, it does not seem necessary to simulate the entire array. While the limitation of computer systems forces a practical limit, so far it does not appear that the modeling methodology itself is dimension-limited.

Software availability. The *R* code used to perform the various steps in this method is available from the authors. The modified Cholesky factorization algorithm as implemented in FORTRAN (Schnabel and Eskow, 1999) was called from the R script.

Acknowledgement

The authors would like to thank Sun Microsystems Computer Company for a grant of high-performance computing equipment which facilitated this work.

References

Affymetrix, , 2002. Statistical Algorithms Description Document. Affymetrix, Inc., Santa Clara, CA. http://www.affymetrix.com/support/technical/whitepapers/sadd-whitepaper.pdf.

Albers, C.J., Jansen, R.C., Kok, J., Kuipers, O.P., van Hijum, S., 2006. SIMAGE: Simulation of DNA-microarray gene expression data. BMC Bioinformatics 7, 205.

Allison, D.B., Gadbury, G.L., Heo, M., Fernandez, J.R., Lee, C-K., Prolla, A., Weindruch, R., 2002. A mixture model approach for the analysis of microarray gene expression data. Computational Statistics and Data Analysis 39, 1–20.

Bioconductor, 2003. Methods for Affymetrix Oligonucleotide Arrays (affy), Bioconductor Project, Version 1.2. http://www.bioconductor.org/.

Bowman, K.O., Shenton, L.R., 1989. S_B and S_U distributions fitted by percentiles: A general criterion. Communications in Statistics, Part B – Simulation and Computation 18, 1–13.

Box, G.E.P., Cox, D.R., 1964. The analysis of transformations (with discussion). Journal of the Royal Statistical Society B 26, 211–252.

Giles, P.J., Kipling, D., 2003. Normality of oligonucleotide microarray data and implications for parametric statistical analyses. Bioinformatics 19 (17), 2254–2262.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, pp. 502–504.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., Speed, T.P., 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4, 249–264.

Johnson, M.E., 1987. Multivariate Statistical Simulation: A Guide to Selecting and Generating Continuous Multivariate Distributions. John Wiley & Sons, New York.

Johnson, N.L., 1949. Systems of frequency curves generated by methods of translation. Biometrika 36, 149–176.

Johnson, N.L., 1965. Tables to facilitate fitting S_U frequency curves. Biometrika 52, 547–558.

Johnson, N.L., Kitchen, J.O., 1971. Some notes on tables to facilitate fitting S_B curves. Biometrika 58, 223–226.

Johnson, N.L., Kotz, S., 1970. Distributions in Statistics: Continuous Univariate Distributions, vol. 1. Houghton-Mifflin Company, Boston.

Johnson, N.L., Kotz, S., 1972. Distributions in Statistics: Continuous Multivariate Distributions. John Wiley and Sons, New York.

Kennedy, W.J., Gentle, J.E., 1980. Statistical Computing. Marcel-Dekker, New York.

Kirisci, L., Al-Subaihi, A.A., Tarter, R., 2005. Effects of the generalized Box–Cox transformation on the type I error rate and power of Hotelling's T^2. Journal of Statistical Computation and Simulation 75 (3), 199–206.

Li, C., Wong, W.H., 2001. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. Proceedings of the National Academy of Sciences USA 98, 31–36.

Molinaro, A.M., Simon, R., Pfeiffer, R.M., 2005. Prediction error estimation: A comparison of resampling methods. Bioinformatics 21 (15), 3301–3307.

Morrison, D.F., 1976. Multivariate Statistical Methods, 2nd edition. McGraw-Hill, New York.

Nguyen, D.V., Arpat, A.B., Wang, N., Carroll, R.J., 2002. DNA microarray experiments: Biological and technological aspects. Biometrics 58, 701–717.

Nykter, M., Aho, T., Ahdesmaki, P., Lehmussola, A., Yli-Harja, O., 2006. Simulation of microarray data with realistic characteristics. BMC Bioinformatics 7, 349.

Rose, C., Smith, M.D., 2002. Mathematical Statistics with Mathematica. Springer, New York.

Sakia, R.M., 1992. The Box–Cox transformation technique: A review. The Statistician 41, 169–178.

Shayib, M.A., 1989. The procedure for selection of transformations from the Johnson system. Communications in Statistics, Part B – Simulation and Computation 18, 1457–1464.

Singh, D., Febbo, P.G., Ross, K., et al., 2002. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1, 203–209.

Singhal, S., Kyvernitis, C.G., Johnson, S.W., Kaisera, L.R., Liebman, M.N., Albelda, S.M., 2003. Microarray data simulator for improved selection of differentially expressed genes. Cancer Biology and Therapeutics 2, 383–391.

Slifker, J.F., Shapiro, S.S., 1980. The Johnson system: Selection and parameter estimation. Technometrics 22, 239–246.

Schnabel, R.B., Eskow, E., 1990. A new modified Cholesky factorization. SIAM Journal of Scientific Statistical Computing 11, 1136–1158.

Schnabel, R.B., Eskow, E., 1991. Algorithm 695: Software for a new modified Cholesky factorization. Transactions on Mathematical Software 17 (3), 306–312.

Schnabel, R.B., Eskow, E., 1999. A revised modified Cholesky factorization algorithm. SIAM Journal on Optimization 9 (4), 1135–1148.

Wheeler, R.E., 1980. Quantile estimators of Johnson curve parameters. Biometrika 67, 725–728.

Xu, P., Brock, G.N., Parrish, R.S., 2009. Modified linear discriminant analysis approaches for classification of high-dimensional microarray data. Computational Statistics and Data Analysis 53 (5), 1674–1687.

Zhou, B., McTague, J.P., 1996. Comparison and evaluation of five methods of estimation of the Johnson system parameters. Canadian Journal of Forest Research 26, 928–935.