University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2008 *Paper* 238

Supervised Distance Matrices: Theory and Applications to Genomics

Katherine S. POLLARD* Mark J. van der Laan†

*UC Davis Genome Center & Dept. of StatisticS, kspollard@ucdavis.edu

†University of California - Berkeley, laan@berkeley.edu

Supervised Distance Matrices: Theory and Applications to Genomics

Katherine S. POLLARD and Mark J. van der Laan

Abstract

We propose a new approach to studying the relationship between a very high dimensional random variable and an outcome. Our method is based on a novel concept, the supervised distance matrix, which quantifies pairwise similarity between variables based on their association with the outcome. A supervised distance matrix is derived in two stages. The first stage involves a transformation based on a particular model for association. In particular, one might regress the outcome on each variable and then use the residuals or the influence curve from each regression as a data transformation. In the second stage, a choice of distance measure is used to compute all pairwise distances between variables in this transformed data. When the outcome is right-censored, we show that the supervised distance matrix can be consistently estimated using inverse probability of censoring weighted (IPCW) estimators based on the mean and covariance of the transformed data. The proposed methodology is illustrated with examples of gene expression data analysis with a survival outcome. This approach is widely applicable in genomics and other fields where high-dimensional data is collected on each subject.

1. Introduction

Due to various technological advances, it is now common to collect very high dimensional data on each subject in a study. We will focus in this paper on gene expression data, although similar data structures arise in proteomics, metabolomics, and many fields outside genomics. A typical microarray experiment results in an observed data matrix X whose columns are n i.i.d. copies of a p-dimensional vector of gene expression measurements. In addition to measuring gene expression, researchers – particularly in clinical settings – are now collecting covariate and outcome data on each sample. With this data, we can extend exploratory methods for finding patterns in gene expression data and begin to study the relationships between gene expression and end points of interest, such as tumor grade, time to metastasis, or survival in cancer patients. Such studies provide insight into disease mechanism. Discovering groups of genes with similar relationships to an outcome is also an important step in designing molecular diagnostic tools.

Gene expression profiling has become an established method for classifying patients into different disease subpopulations. Associations between messenger RNA (mRNA) expression signatures and clinical outcomes have been discovered in several studies (*e.g.* Rosenwald et al.(2003)Rosenwald, Wright, Wiestner, Chan, Conors et al.). Even more striking associations with disease subtypes have been discovered for microRNA (miRNA) expression profiles, which can now be measured in a similar high-throughput manner (Lu et al.(2005)Lu, Getz, Miska, Alvarez-Saavedra, Lamb et al.). One goal of such studies is to develop molecular signatures that can be used to better diagnose and tailor treatment for future patients. Because of the large number of genes assayed in a microarray experiment, serious attention has been devoted to the issue of dimension reduction in prediction problems using gene expresesion data (*e.g.* Li & Li(2004); Nguyen & Rocke(2002)). It is often the case that many genes are more or less equally predictive of the outcome of interest and that colinearity between genes makes variable selection unstable between repeated experiments. This suggests that it would be useful to identify groups of genes whose expression profiles have a similar association with the outcome variable.

Supervised clustering methods aim to group genes based on the association between their expression profile across subjects and a supervising variable measured on the same subjects. The supervision can simply be based on a set of known or pre-defined expression profiles, in which

case the goal is to find genes that match each profile (*e.g.* Qu & Xu(2004)). More generally, the idea is to put genes together that have a similar relationship to a variable of interest, regardless of whether or not they have similar expression profiles. This latter approach has been implemented for a binary outcome using support vector machines (Brown et al.(2000)Brown, Grundy, Lin, Cristianini, Sugnet et al.), for categorical outcomes using a forward-backward search algorithm (Dettling & Bühlmann(2002)), and for contnuous outcomes using gene shaving (Hastie et al.(2000)Hastie, Tibshirani, Eisen, Alizadeh, Levy et al.). Each of these methods is based on a particular choice of clustering algorithm, including a way to measure distance between genes and a criteria for quantifying cluster homogeneity.

In this paper, we propose a general approach to supervised clustering that can be used with any choice of distance and clustering algorithm. Our main contribution is the idea of a supervised distance matrix, which measures similarity between variables (*e.g.* genes) based on their association with an outcome. We show that the supervised distance matrix can be consistently estimated even when the outcome is right-censored. As an illustration of the methodology, we focus on understanding the association of gene expression and a post-expression outcome such as survival.

2. Data and Notation

Consider a p-dimensional random vector X and a univariate random variable Y. To be concrete, we will talk about gene expression data, where X is a vector of expression levels for p genes and Y is an outcome of interest, which may be right-censored. When Y is censored, we do not observe the full data (X, Y), but rather $O = (Y \wedge C, \Delta = I[Y \leq C], X)$, where C is the censoring time. If additional covariates are measured, we denote these by V and then $O = (Y \wedge C, \Delta = I[Y \leq C], X, V)$. Suppose we observe a sample of n i.i.d. copies of O. The observations of X can be stored in a $n \times p$ matrix $\mathbf{X} = \mathbf{X}(i, j)$ whose i'th row is the p-dimensional gene expression profile for subject i and whose j'th column is the vector of n gene expression values for gene j across subjects. As short-hand, let X_j denote the j'th column of $\mathbf{X}$.

3. Supervised Distance Matrices

Our goal is to define a measure of pairwise distance that reflects the degree to which the expression data X_j and $X_{j'}$ for genes j and j' ($j = 1, \ldots, p$, $j' = 1, \ldots, p$) have similar patterns of covariation with the outcome Y across subjects. Our approach to defining such a supervised distance matrix is first to transform the data $(\mathbf{X}, Y)$ to form a matrix of "association profiles" and then to compute pairwise distances between these transformed data profiles.

3.1 *Stage 1: Transformations*

Let $\mathbf{W} = \mathbf{W}(\mathbf{X}, Y)$ be a transformation of the multivariate data $\mathbf{X}$ (*i.e.* gene expression profiles) and the outcome Y based on a choice of model(s) for the marginal association of X_j with Y, $j = 1, \ldots, p$. In particular, we might choose a common regression model $m(x \mid \beta)$ for all genes, parametrized by a finite dimensional parameter β, so that $E(Y \mid X_j) = m(X_j \mid \beta_j)$, $j = 1, \ldots, p$. Then, we can define an association profile $W_j = W(\mathbf{X}, Y|\beta_j)$ for each gene. The association profile W_j will typically be n-dimensional, representing random deviations from an average (across the whole population) association of X_j with Y as quantified by the particular gene-specific regression model. In this case, it makes sense to store these profiles in an $n \times p$ matrix $\mathbf{W}$ of the same dimension as $\mathbf{X}$. We can think of each row of $\mathbf{W}$ as a realization of a random variable W, which is a p-dimensional vector $(\mathbf{W}(i,1), \ldots, \mathbf{W}(i,p))$ whose j-th component represents the subject-specific association of gene j's expression with the outcome Y for subject i ($i = 1, \ldots, n$, $j = 1, \ldots, p$).

Table 1 gives several specific examples of transformations W_j that provide meaningful association profiles for the linear regression model

$$Y = m(X_j \mid \beta_j) + \epsilon_j = \beta_{j0} + \beta_{j1} X_j + \epsilon_j$$

The transformations are defined as if the regression parameters $\beta = (\beta_{j0}, \beta_{j1} : j = 1, \ldots, p)$ were known. Estimation is discussed in Section 4. The same approach can be used to define transformations for other models.

[Table 1 about here.]

- **Regression Coefficients.** A simple transformation is the regression coefficient $W_j(\mathbf{X}, Y) = \beta_{j1}$ ($j = 1, \ldots, p$) for the slope. In this case, W_j is a single number, not an n-vector. Like

means in a typical gene expression analysis, marginal regression coefficients can be useful for gene selection, but they are not useful for clustering.

- **Residuals.** The vector of residuals $\epsilon_j = Y - m(X_j|\beta_j)$ captures subject-specific deviations from $E(Y \mid X_j)$. The transformation $W_j(\mathbf{X}, Y) = \epsilon_j$ is intuitively appealing, since residuals are widely used in statistics and biostatistics for assessing subject-specific contributions and goodness of fit in regression model diagnostics.

- **Influence Functions.** The influence function $\mathbf{IC}(i, j)$ represents the contribution of subject i to the regression of gene j on Y. As such, $W_j(\mathbf{X}, Y) = (\mathbf{IC}(1, j), \ldots, \mathbf{IC}(n, j))$ is an interesting association profile. In the case of the linear regression model with the intercept β_{j0} known, the efficient influence curve for the slope parameter β_{j1} can be thought of as a subject-specific deviation from the overall slope (Appendix A).

- **Standardized Residuals.** Another transformation is the vector of standardized residuals $W_j(\mathbf{X}, Y) = (\epsilon(1, j)/\mathbf{X}(1, j), \ldots, \epsilon(n, j)/\mathbf{X}(n, j))$, where $\epsilon = \epsilon(i, j)$ is the matrix of residuals for subject i and gene j. A connection between the efficient influence curve and the transformation ϵ_j/X_j is given in Appendix A.

Remark 1: The transformation $W_j = \epsilon_j/X_j$ will be unstable at small values of X_j. We propose, therefore, the transformation $\epsilon_j/\tilde{X}_j = \frac{\epsilon_j}{X_j+\delta}$, where δ is a data adaptively selected small number added to X_j for robustness against very small X_j. Alternatively, one could use $max(X_j, c)$ in the denominator to truncate gene expression from below by a constant c.

Remark 2: For each of these transformations, it is also possible to include covariates in the model for association. The residuals from $Y = m(X_j, V \mid \beta_j) + \epsilon_j$ are adjusted for the covariate(s) V. In this case, the transformation is $W_j = W_j(V, \mathbf{X}, Y)$, denoting the dependence on V.

Remark 3: For generalized linear regression, influence functions are equivalent to subject-specific deviations from the overall regression coefficient to a first order approximation. The efficient influence curve suggests a transformation $W_j = \epsilon_j \left\{ \frac{d}{d\beta_{j1}} m(X_j|\beta) \right\}^{-1}$ (Appendix B).

3.2 *Stage 2: Distances*

Given a transformation $\mathbf{W} = \mathbf{W}(\mathbf{X}, Y)$ of the gene expression data and outcome into n-dimensional association profiles $\{W_j : j = 1, \ldots, p\}$, a $p \times p$ empirical supervised distance matrix $\tilde{\mathbf{D}}$ is obtained by simply applying a choice of pair-wise distance (metric or non-metric) to the columns of $\mathbf{W}$. Some examples include Euclidean, cosine-angle, and correlation distance. For a given choice of distance d, $\tilde{\mathbf{D}} = d(\mathbf{W}) = \{d(W_j, W_{j'}) : j = 1, \ldots, p, j' = 1, \ldots, p\}$ measures dissimilarity between pairs of gene association profiles. In other words, the distance $\tilde{\mathbf{D}}(j, j') = d(W_j, W_{j'})$ is small if genes j and j' have a similar association between expression and the outcome Y across the n subjects. Because $\tilde{\mathbf{D}}$ is based on association profiles $\mathbf{W}$, it directly reflects distance between genes based on their associations with Y, rather than their expression per se.

We refer to the matrix $\tilde{\mathbf{D}} = d(\mathbf{W})$ as an empirical distance matrix, because it is based on the transformed data $\mathbf{W}$ for a sample of size n. In the discussion above, we suppose that the true transformation $\mathbf{W}$ is known. In practice, $\mathbf{W}$ and hence the empirical distance matrix $\tilde{\mathbf{D}}$ must be estimated. Estimation of the transformations $W_j = W_j(\mathbf{X}, Y)$ involves fitting an appropriate regression model $E(Y|X_j) = m(X_j \mid \beta_j)$ for the association between the expression profile of each gene X_j and the outcome Y. The estimator $\hat{\beta}_j$ provides an estimator $\hat{W}_j = W_j(\mathbf{X}, Y|\hat{\beta}_j)$ of W_j. For instance, the transformations W_j in Table 1 are based on the intercept and slope parameters $\beta_j = (\beta_{j0}, \beta_{j1})$ from a simple linear regression model. An estimator $\hat{\beta}_j$ can be obtained through maximum likelihood (or least squares). Let $\hat{\mathbf{D}} = d(\hat{\mathbf{W}}) = \{d(\hat{W}_j, \hat{W}_{j'}) : j = 1, \ldots, p, j' = 1, \ldots, p\}$ denote the estimated supervised distance matrix, which is an empirical supervised distance matrix based on the estimated transformation $\hat{\mathbf{W}}$.

3.3 *Clustering*

We now make a few comments about the use of the proposed empirical supervised distance matrices to cluster genes with regard to their association profiles. Recall that in unsupervised clustering of gene expression data $\mathbf{X}$, the distance matrix measures pair-wise distances between the genes' expression profiles $\{X_j : j = 1, \ldots, p\}$, and the goal of clustering is to find groups of genes whose expression profiles X_j are similar. Here, we propose to supervise the clustering of gene expression profiles with the outcome of interest Y by using instead the estimated supervised distance matrix $\hat{\mathbf{D}}$. Thus, standard clustering methods can be applied directly to the analysis of

the association between $\mathbf{X}$ and Y by using the matrix $\hat{\mathbf{D}}$ as input. In particular, any unsupervised clustering algorithm can now be employed for supervised clustering.

A strong cluster of genes in $\hat{\mathbf{D}}$ represents a group of genes which show the same association between Y (*e.g.*: survival) and gene expression across subjects. The n-dimensional profile of this cluster, such as a cluster mean or medoid, identifies the typical response of Y to these genes. The pattern of this response can vary between gene clusters. For example, for a given cluster we might find that either all subjects show the same response to these genes, or the subjects cluster into two or more groups with respect to these genes, or the subjects might show a gradient of increasing responses.

Remark 4: Typically, the dimension of a gene expression data set is reduced before clustering by removing any genes that do not carry significant information about the question of interest. Filtering rules are usually based on testing a null hypothesis for each gene and making rejection decisions so that a multiple testing error rate is controlled. The same methods that are employed for filtering the gene expression profiles $\mathbf{X}$ can be applied to the transformed data $\hat{\mathbf{W}}$. For example, testing the null hypotheses $H_0(j) : \beta_{j1} = 0$ allows one to remove genes whose profiles show no marginal association with the outcome Y.

4. Consistency Theorems for Supervised Distance Matrices

In the previous section, we focused on how one might use estimated supervised distance matrices in practice. We now make the observation that there exists some true supervised distance matrix $\mathbf{D}$, which can be thought of as a parameter of the data generating distribution. This matrix $\mathbf{D} = \mathbf{D}(j, j')$ measures how similar genes j and j' are in terms of their associations with the outcome Y in the population. We can therefore think of the empirical distance matrix $\tilde{\mathbf{D}}$ as an empirical estimator of $\mathbf{D}$ based on the true parameters β_j. Similarly, $\hat{\mathbf{D}}$ is an empirical estimator of $\mathbf{D}$ based on estimated parameters $\hat{\beta}_j$. Note that $\hat{\mathbf{D}}$ is the estimator one would typically use in practice, since the regression parameters will not usually be known. Both estimators rely on computing an empirical distance from a (possibly estimated) transformation of the observed data. In this section, we show that under certain conditions on the data and the transformation, $\mathbf{D}$ is

consistently estimated whenever $\hat{\beta}_j$ is a consistent estimator of β_j.

4.1 *Uncensored Data*

If the outcome Y is observed for all subjects, consistent estimation of the supervised distance matrix $\mathbf{D}$ amounts to consistent estimation of the transformation $\mathbf{W}$, since $\mathbf{D}$ is a deterministic function of $\mathbf{W}$. First, consider the specific case where the transformation W_j is the n-vector of linear regression residuals $\epsilon_j = Y - (\beta_{j0} + \beta_{j1} X_j)$ for gene j. If X_j is bounded and $\hat{\beta}_j$ converges to β_j in probability as $\frac{n}{\log p} \to \infty$, then the following theorem shows that $\hat{\mathbf{D}}$ converges to $\mathbf{D}$ at the same rate.

THEOREM 1. *Let $W_j = Y - (\beta_{j0} + \beta_{j1} X_j)$, where Y is not censored. Suppose $|Y| \leq M$ and $|X_j| \leq M$ for a constant $M > 0$, $j = 1, \ldots, p$. If $\sup_j |\hat{\beta}_j - \beta_j| \xrightarrow[\frac{n}{\log p} \to \infty]{P} 0$, then*

$$\sup_{j,j'} |\hat{\mathbf{D}}(j, j') - \mathbf{D}(j, j')| \xrightarrow[\frac{n}{\log p} \to \infty]{P} 0.$$

The proof of Theorem 1 is given in Appendix C. It involves showing that $\hat{\mathbf{W}}$ converges to $\mathbf{W}$ as $\frac{n}{\log p} \to \infty$, using Berstein's inequality. Then the result follows, since $\mathbf{D}$ is a deterministic function of $\mathbf{W}$ and $\hat{\mathbf{D}}$ is a deterministic function of $\hat{\mathbf{W}}$.

Next, consider a general transformation W_j. Under the conditions of the following theorem, similar reasoning to that used for the residual transformation provides convergence of $\hat{\mathbf{D}}$ to $\mathbf{D}$.

THEOREM 2. *Let $W_j = W_j(\beta_j)$ be a transformation of the data $(\mathbf{X}, Y)$ that is a function of an unknown regression parameter β_j, $j = 1, \ldots, p$. Suppose Y is not censored. Consider the estimator $\hat{W}_j = W_j(\hat{\beta}_j)$. If $|Y| \leq M$ and $W_j(\hat{\beta}_j) - W_j(\beta_j) \leq M(\hat{\beta}_j - \beta_j)$ for a constant $M > 0$ and if $\sup_j |\hat{\beta}_j - \beta_j| \xrightarrow[\frac{n}{\log p} \to \infty]{P} 0$, then*

$$\sup_{j,j'} |\hat{\mathbf{D}}(j, j') - \mathbf{D}(j, j')| \xrightarrow[\frac{n}{\log p} \to \infty]{P} 0.$$

The proof of Theorem 2 follows the same Bernstein's inequality argument as the proof of Theorem 1, with the condition $W_j(\hat{\beta}_j) - W_j(\beta_j) \leq M(\hat{\beta}_j - \beta_j)$ playing the role of $|X_j| \leq M$.

We now make an observation that will allow us to estimate supervised distance matrices even when the outcome Y is censored. Note that we have been describing the supervised distance

matrix $\mathbf{D}$ as a deterministic function of the transformed data $\mathbf{W}$. In fact, $\mathbf{D}$ is typically also a deterministic function of the mean $\mu = E(\mathbf{W})$ and covariance $\Sigma = E(\mathbf{W} - \mu)(\mathbf{W} - \mu)^\top$ of $\mathbf{W}$, which are parameters of the underlying data generating distribution. This is the case for many commonly employed distance metrics, including Euclidean, cosine-angle, and correlation distance (as well as the absolute values of these). For example, the Euclidean distance matrix is given by:

$$d(W_j, W_{j'}) = n(\sigma_{jj} + \sigma_{j'j'} - 2\sigma_{jj'} + (\mu_j - \mu_{j'})^2). \tag{1}$$

When $\mathbf{D} = \mathbf{D}(\mu, \Sigma)$ is a deterministic function of the mean and covariance of $\mathbf{W}$, we can use the estimator $\bar{\mathbf{D}} = \mathbf{D}(\hat{\mu}, \hat{\Sigma})$ based on estimates of the mean and covariance of $\mathbf{W}$. The following theorem is the analog of Theorem 1 for convergence of $\bar{\mathbf{D}}$.

THEOREM 3. *Let $W_j = Y - (\beta_{j0} + \beta_{j1} X_j)$, where Y is not censored. Consider the estimator $\bar{\mathbf{D}} = \mathbf{D}(\hat{\mu}, \hat{\Sigma})$ of $\mathbf{D} = \mathbf{D}(\mu, \Sigma)$, defined above. Suppose $|Y| \leq M$, $|X_j| \leq M$, $|W_j| \leq M$, $|\beta_j| \leq M$, and $|\hat{\beta}_j| \leq M$ for a constant $M > 0$, $j = 1, \ldots, p$. If the variance of X_j is bounded away from zero uniformly in j, then*

$$\sup_{j,j'} |\bar{\mathbf{D}}(jj') - \mathbf{D}(jj')| \xrightarrow[\frac{n}{\log p} \to \infty]{P} 0.$$

The proof, given in Appendix C, involves showing that the first two moments of the distribution of W_j, (μ, Σ), can be consistently estimated under the conditions of the theorem. A similar result can be obtained for other transformations.

4.2 *Censored Data*

For uncensored data, it is not necessary to use the estimator $\bar{\mathbf{D}} = \mathbf{D}(\hat{\mu}, \hat{\Sigma})$ of the supervised distance matrix, since we can typically estimate the transformation $\mathbf{W}$ itself and use $\hat{\mathbf{D}} = d(\hat{\mathbf{W}})$. However, when the transformation is not directly estimable, due to Y being unobserved for some subjects, this alternative estimator $\bar{\mathbf{D}}(\hat{\mu}, \hat{\Sigma})$ provides an approach to compute supervised distance matrices in the presence of censoring.

A method for the estimation of regression coefficients when the outcome is right-censored is based on the use of Horvitz-Thompson type estimators called Inverse Probability of Censoring Weighted (IPCW) estimators, which are presented in great generality in Robins & Rotnitzky(1992)). The optimal estimating function is discussed in Robins & Rotnitzky(1992)) and

van der Laan & Robins(2003)) (Chapter 3), but optimality is not the focus of this paper. Here, we aim to illustrate how one example of a simple IPCW estimator can be used in the estimation of the mean and covariance of $\mathbf{W}$ when Y is censored. A similar approach can be used with any such estimator. The idea behind IPCW estimators is to use the data from uncensored subjects to estimate the mean and covariance (μ, Σ) of $\mathbf{W}$. The estimators are weighted by the inverse probability of censoring given the data on a subject. These weights make the estimator unbiased.

The IPCW estimators of μ and $\Sigma = \{\sigma_{jj'}\}$ based on the observed data $O_i = (Y_i \wedge C_i, \Delta_i = I[Y_i \leq C_i], X_i), i = 1, \ldots, n$ are given by:

$$\hat{\mu}_j = \hat{E}(\hat{W}_j) = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{\mathbf{W}}(i,j)\Delta_i}{\bar{G}_n(Y_i \mid \mathbf{X}, Y)}, j = 1, \ldots, p, \tag{2}$$

$$\hat{\sigma}_{jj',n} = \hat{E}(\hat{W}_j \hat{W}_{j'}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{\mathbf{W}}(i,j)\hat{\mathbf{W}}(i,j')\Delta_i}{\bar{G}_n(Y_i \mid \mathbf{X}, Y)}, j = 1, \ldots, p, j' = 1, \ldots, p, \tag{3}$$

where $\bar{G}(t \mid \mathbf{X}, Y) = pr(C_i > t \mid \mathbf{X}, Y)$ is the probability that subject i was still at risk at time t given his/her gene expression profile. We call $\bar{G}$ the censoring mechanism. $\bar{G}(t \mid \mathbf{X}, Y)$ is estimated by $\bar{G}_n(t \mid \mathbf{X}, Y)$. Note that the proposed IPCW estimators involve two estimation steps: estimation of the transformation W_j for uncensored subjects and estimation of the mean or covariance via the empirical mean and covariance. If the true transformation were know, we could form estimates $\tilde{\mu}_{j,n}$ and $\tilde{\Sigma}_{jj',n}$ similar to Equations 2 and 3 (respectively) by replacing $\hat{\mathbf{W}}(i,j)$ with $\mathbf{W}(i,j)$ and $\bar{G}_n(Y_i \mid \mathbf{X}, Y)$ with $\bar{G}(Y_i \mid \mathbf{X}, Y)$ in each expression.

Given a choice of distance, the $p \times p$ supervised distance matrix $\mathbf{D}$ is estimated by plugging in the IPCW estimators $(\hat{\mu}, \hat{\Sigma})$ to Equation 1 or its analog. The resulting IPCW supervised distance matrix estimator $\bar{\mathbf{D}}$ can then be used for clustering as described above for uncensored data. We now turn to the question of consistency. Convergence of $\bar{\mathbf{D}} = \mathbf{D}(\hat{\mu}, \hat{\Sigma})$ to $\mathbf{D}$ depends on consistency of the estimators $(\hat{\mu}, \hat{\Sigma})$, since $\mathbf{D}$ is a deterministic function of (μ, Σ). The following theorem gives the necessary conditions for convergence of the IPCW estimators (Equations 2 and 3) to (μ, Σ) for the residual transformation.

THEOREM 4. *Let $W_j = Y - (\beta_{j0} + \beta_{j1}X_j)$, where Y is right-censored. Consider the estimator $\bar{\mathbf{D}} = \mathbf{D}(\hat{\mu}, \hat{\Sigma})$ of $\mathbf{D} = \mathbf{D}(\mu, \Sigma)$, defined above. Suppose $|Y| \leq M$, $|X_j| \leq M$, $|W_j| \leq M$, $|\beta_j| \leq M$, and $|\hat{\beta}_j| \leq M$ for a constant $M > 0$, $j = 1, \ldots, p$. Also suppose the variance of X_j is bounded away*

from zero uniformly in j. Assume that (i) $C \perp Y|X$, (ii) $\sup|\bar{G}_n(Y|\mathbf{X},Y) - \bar{G}(Y|\mathbf{X},Y)| \xrightarrow{P} 0$ *where the supremum is over the support of the distribution of* $(\mathbf{X},Y)$*, and (iii)* $\bar{G}(Y|\mathbf{X},Y) > \delta > 0$ *for a.e.* $(\mathbf{X},Y)$*, where* $\bar{G}(t|\mathbf{X},Y) = Pr(C > t|\mathbf{X},Y)$*. Then,*

$$\sup_{j,j'} |\bar{\mathbf{D}}(j,j') - \mathbf{D}(j,j')| \xrightarrow[\frac{n}{\log p}\to\infty]{P} 0.$$

The proof is given in Appendix D. Similar results can be obtained for other transformations.

Remark 5: One can think of the output of a clustering algorithm (*e.g.* gene cluster labels or a hierarchical tree) as parameters of the underlying data generating distribution. In the supervised clustering problem presented here, these clustering parameters are typically deterministic functions of the supervised distance matrix $\mathbf{D}$. Hence, we can consistently estimate the supervised clustering parameters themselves as long as we can consistently estimate $\mathbf{D}$.

5. Simulations

In order to illustrate the implementation of this method, we designed a simulation consisting of gene expression and survival time (possibly censored) for a sample of subjects. We generated the data in such a way that one gene (the causal gene g_1) is perfectly predictive of survival and another nine genes have expression *very* similar to this gene. The remaining genes have one of several expression patterns, one of which has the same mean as the causal gene. Thus, the genes still form clusters with respect to gene expression alone, but an interesting cluster of ten genes exists which has a special relationship to survival.

5.1 *Data Generation*

First, we generate the gene expression matrix X with $n = 30$ patients and $p = 1010$ genes. The effect of increasing the sample size to $n = 100$ is investigated later. We suppose that the genes with insignificant difference in expression between tumor and healthy tissues have already been removed from the data set. For 1000 genes, each gene's expression is an independent $N(m, 0.75)$ variable, where $m \in \{-9, -8, -5, -4, 4, 5, 8, 9\}$ and each mean group consists of 125 genes. In addition, we generate 10 genes with $m = -9$, that are not independent. A single gene g_1 is generated first, and then the other nine genes are g_1 plus random $N(0, 0.05)$ noise. In this way, we have an extra ten genes with mean $m = -9$ that are very close to each other.

Next, we generate survival times for each patient as a deterministic function of the expression of g_1. We generate according to $\log T_i = \beta_{0i} + \beta_{1i} g_{1i}$. For patients 1 to 15, we set $\beta_{0i} = 0$, while for patients 16 to 30 we set $\beta_{0i} = -2$. For all patients, we set $\beta_{1i} = -1$. Finally, we generate a censoring time for each patient. We consider no censoring, 20%, and 30% expected censoring. Since the maximum log survival time is ≈ 10, we generate the censoring times $\log C$ from $U(0, 10/q))$, where q is the expected fraction of patients we wish to have censored.

5.2 *Method*

We use the clustering algorithm PAM (Kaufman & Rousseeuw(1990)) throughout the simulations. The emphasis in this paper is not on the choice of algorithm, but rather on the transformation method. We like PAM for our purposes, because it allows any user-supplied distance metric and the medoids (elements themselves) are robust representations of the clusters. We follow the recommendation of Kaufman & Rousseeuw(1990)) and chose the number of clusters by maximizing average silhouette, a measure of how well matched elements are to their own cluster versus the next closest cluster. We use Euclidean distance, which is capable of detecting differences between groups of genes differing in mean expression.

First, we cluster genes using the gene expression data X only. Next, we fit a linear regression model for $\log T$ and each gene's expression. We look at which genes have t-statistics larger than expected using a simple Bonferoni adjustment. In the uncensored data simulation, we are able to compute the residual transformation $\tilde{X}$ directly. We then calculate the Euclidean distance matrix from $\tilde{X}$ and cluster genes.

In the simulations with censoring, we use IPCW estimators for the mean and covariance of $\tilde{X}$ to calculate the gene Euclidean distance matrix. We use a Cox proportional hazards model for the censoring mechanism. Since we know that none of the gene's are associated with censoring, we choose to use only g_1 to fit the model for $pr(C > t \mid X)$. Gene g_1 is a sensible choice, because its expression is most associated with survival time so that by including it in the model for the censoring mechanism we gain efficiency (van der Laan & Robins(2003)), p.135). We cluster genes using the transformed gene Euclidean distance matrix.

5.3 *Results*

For comparative purposes, we first describe the results of applying standard unsupervised clustering to the simulated data. Then, the results from supervised clsutering are presented.

Gene Expression Only. Average silhouette suggests that there are two clusters in the gene expression data. We apply PAM with $k = 2$ and find that these are the over and under expressed genes (means less than and greater than zero). When we also try $k = 8$ clusters, PAM identifies the eight groups based on means.

Residual Transformation. We apply the residual transformatin approach to simulated data without censoring and with right censoring of 20% or 30% of subjects.

1. Without Censoring:

   - There are two gene clusters, which correspond exactly to g_1's group (C_1) and the rest of the genes (C_2).
   - Figure 1 illustrates the presence of two patient subpopulations in C_1. This separation of the patients into subpopulations is not evident when all genes are used nor when the genes in C_1 are used but the distance matrix is calculated from gene expression alone. This result highlights a situation where we can identify an interesting patient subpopulation which would not have been evident without a sensible transformation.

[Figure 1 about here.]

2. With Censoring (using IPCW estimators):

   - First, we consider C distributed $U(0, 50)$, so that about one fifth of the patients are censored. There are two clusters, which correspond with g_1's group plus nine other genes (C_1) versus the rest of the genes (C_2).
   - With C distributed $U(0, 30)$, the gene distance matrix computed from the IPCW mean and covariance estimates again has two clusters. The cluster with g_1's group now contains 50 genes, indicating that for $n = 30$ and 30% censoring it is harder to estimate the transformed data matrix than with only 20% censoring. Figure 2 shows the two distance matrices.
   - When the number of subjects is increased to $n = 100$, g_1's group is identified exactly as one cluster, even with 30% censoring.

[Figure 2 about here.]

6. Data Analysis

We apply the methodology proposed in this paper to a puclically available data set that includes measures of gene expression and survival for 92 patients with mantle cell lymphoma (MCL), a non-Hodgkin's lymphoma Rosenwald et al.(2003)Rosenwald, Wright, Wiestner, Chan, Conors et al.). All patients were cyclin D1 negative. Sixty-four of the patients died during the course of the study, while the remaining 28 patients were right-censored. Expression data was available for 8810 genes. Based on previous studies and this data, the authors identify a set of genes (the "proliferation signature") that are involved in cell proliferation and are predictive of survival. The data set includes the mean expression profile for this set of genes.

6.1 *Gene Filtering*

We first screen out any genes that do not show an assoication with survival. One might explore a number of gene selection strategies. Here, we follow the simulations and fit a linear model for log survival time as a function of the expression profile of each gene. Then, for each gene, we test the significance of the association between that gene's expression and survival time in the fitted model using a standard t-test. We select for cluster analysis any gene with p-value $p < 0.01$. This produces a set of 750 genes.

One could, of course, fit other models for survival and use alternative filtering procedures. Each choice of model and procedure would produce a potentially different set of genes for clustering. For example, one might choose to use a more non-parametric test that accounts for multiple testing. Several options are implemented in the R `multtest` package available at `http://bioconductor.org` (Pollard & van der Laan(2004) Dudoit et al.(2004)Dudoit, van der Laan & Pollard). Since our goal here is simply to reduce the number of genes for clustering in a straightforward and computationally easy way, the t-test is a reasonable choice.

6.2 *Supervised Clustering*

Next, we compute an IPCW estimator of the Euclidean gene $\times$ gene distance matrix based on the residual transformation. This distance matrix is appropriate for grouping genes based on the mean association between expression level and survival time. In particular, the association profiles for a patient reflect how that patient's genes predict their survival.

Given IPCW estimators $(\hat{\mu}, \hat{\Sigma})$ of the mean and covariance of the transformed data (matrix of residuals), we can compute the estimated supervised Euclidean distance matrix $\bar{\mathbf{D}}$ using Equation 1.

Estimation of $(\hat{\mu}, \hat{\Sigma})$ is based on Equations 2 and 3. We estimate the censoring mechanism, $\bar{G}_n(Y \mid \mathbf{X}, Y)$ using Kaplan-Meier (*i.e.* without using gene expression). A more efficient estimator could be employed if needed. We also explored fitting a Cox proportional hazards model with gene expression as a predictor. Because the gene expression data is so high-dimensional, this involves some model selection or prior knowledge about which genes to include in the model. One sensible option in this data set is to use the mean expression profile of the proliferation signature as predictor. The proliferation signature is a reasonable summary of the full data set $\mathbf{X}$, since the authors identified this variable as predictive of survival. We found that the proliferation signature was only weakly associated with censoring time (coefficient = 0.68, $p = 0.055$). The censoring mechanism based on the estimated survival function from this Cox model is similar to that from Kaplan-Meier. Hence, we used the simpler Kaplan-Meier estimated censoring mechanism to form our estimator $\bar{\mathbf{D}}$.

[Figure 3 about here.]

Any choice of clustering algorithm can now be applied to $\bar{\mathbf{D}}$. Here, we use the same general approach (using the PAM algorithm) that we employed in the simulations. Average silhouette suggests that there are between 2 and 7 clusters (these produce roughly equal values of average silhouette). An alternative criteria, median split silhouette (van der Laan & Pollard(2003))), indicates that there are 7 clusters. Furthermore, results with only a few large clusters are typically difficult to interpret. So, we chose to apply PAM with $k = 7$ clusters. Figure 3 shows $\bar{\mathbf{D}}$ with genes ordered by cluster.

The seven genes in the smallest cluster (cluster 7, the last one in the lower right) are very similar to one another in terms of their association with survival. The residuals for all nine genes show the same gradient across patients. The subset of patients with small residuals represent a sub-population for which these genes are very predictive of survival in the corresponding fitted linear models. This cluster includes a heat shock protein, a splicing factor, a polyprymidine tract binding protein, a zinc finger protein, RNU2, and two hypothetical proteins. It would be interesting to investigate the roles of these genes in MCL. Several other clusters, in particular cluster 6 (also in the lower right), are also fairly striking in terms of the similarity of residual profiles across patients. These clusters provide candidates for studying the coordinated involvement of genes in MCL.

7. Discussion

We have proposed several transformations of a gene expression matrix and an outcome and illustrated that standard clustering methods for gene expression data can be applied to the transformed data matrix in order to discover groups of genes with similar association profiles. This approach can easily adjust for covariates by including these variables in the models for the outcome. Using a simulation, we illustrated the usefulness of the transformation method in a case where two subpopulations have the same gene expression profile for a set of genes, while this set of genes has a different relationship to the outcome in each subpopulation. Therefore, clustering based on distances between gene expression is simply the wrong distance for the purpose of finding such subpopulations.

We have also presented a method for IPCW estimation of the supervised distance matrix and associated clustering parameters when the outcome is right censored. In the simulation, we found that even with sample sizes as small as $n = 30$, we can identify interesting clusters of genes with the IPCW estimators with reasonable amounts of censoring. With more censoring, the genes of interest are still identified, but there are other genes in their cluster as well. Increasing the sample size to $n = 100$ results in these extra genes no longer being clustered with the causal gene, even with as much as 30% censoring. This finding illustrates how simulations can be used to investigate the asymptotic behavior of transformed data clustering parameters (with and without censoring). It is important to understand the true parameters ($n \to \infty$) separately from the problem of estimation in a finite sample.

This IPCW methodology was illustrated on an MCL gene expression data set. The proposed approach allowed us to cluster genes based on their association with survival, even when nearly a third of patients were right-censored. We identified seven distinct groups of genes based on their association profiles. Several of these clusters contain genes with very similar residuals in a model of log survival time as a function of gene expression. In studies with additional clinical information on patients, these clinical variables could be included in the model for survival, producing a supervised distance matrix based on adjusted gene expression association profiles.

Often when one conducts a gene expression study, the goal is to discover underlying causal

relationships and thereby learn transcriptional networks. The method we have presented identifies genes with similar association profiles. The usual caveats about association not implying causation apply in this setting. Investigation of this approach with causal models is a topic for future research.

We have focused on gene expression as a predictor of an outcome, such as survival. The methods we propose can also be applied to study gene expression as an outcome with a treatment or time variable as predictor. In this case, the roles of Y and X_j are reversed in the linear model. Then, the supervised distance matrix $\mathbf{D}$ based on the residual transformation and correlation distance is equivalent to the partial correlation between X_j and $X_{j'}$ adjusting for Y.

While the emphasis in this paper has been on estimation of clustering parameters, it is important to also estimate the variability of these parameter estimates. We previously proposed a statistical framework for analysis of gene expression data and suggested bootstrap methods for statistical inference in this setting (van der Laan & Bryan(2001); Pollard & van der Laan(2002); van der Laan & Pollard(2003)). Since the transformations presented in this paper are deterministic functions of the data generating distribution, this framework for clustering a gene expression matrix can be applied directly to the transformed matrices. The ability to assess reliability is particularly crucial with the high dimensional data structures and relatively small samples in gene expression experiments.

References

Bickel, P., Klaassen, C. J., Ritov, Y. & Wellner, J. (1993). *Efficient and adaptive estimation for semiparametric models.* Baltimore, MD: Johns Hopkins University Press.

Brown, M., Grundy, W., Lin, D., Cristianini, N., Sugnet, C. et al. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Procedings of the National Academy of Sciences* 97 262–267.

Bryan, J. (2001). *Statistical methods for gene expression analysis from cDNA microarrays.* Ph.D. thesis, Division of Biostatistics, University of California, Berkeley.

Dettling, M. & Bühlmann, P. (2002). Supervised clustering of genes. *Genome Biology* 3 1–15.

Dudoit, S., van der Laan, M. J. & Pollard, K. S. (2004). Multiple testing. Part I. Single-step procedures for control of general Type I error rates. *Statistical Applications in Genetics*

and Molecular Biology 3 Article 13.

Hastie, T., Tibshirani, R., Eisen, M., Alizadeh, A., Levy, R. et al. (2000). 'gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology* 1 1–12.

Kaufman, L. & Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons.

Li, L. & Li, H. (2004). Dimension reduction methods for microarrays with application to censored survival data. *Bioinformatics* 20 3406–3412.

Lu, J., Getz, G., Miska, E., Alvarez-Saavedra, E., Lamb, J. et al. (2005). Microrna expression profiles classify human cancers. *Nature* 435 834–838.

Nguyen, D. & Rocke, D. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* 18 39–50.

Pollard, K. & van der Laan, M. (2002). Statistical inference for simultaneous clustering of gene expression data. *Mathematical Biosciences* 176 99–121.

Pollard, K. S. & van der Laan, M. J. (2004). Choice of a null distribution in resampling-based multiple testing. *Journal of Statistical Planning and Inference* 125 85–100.

Qu, Y. & Xu, S. (2004). Supervised cluster analysis for microarray data based on multivariate gaussian mixture. *Bioinformatics* 20 1905–1913.

Robins, J. & Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology, Methodological issues*. Birkhäuser.

Rosenwald, A., Wright, G., Wiestner, A., Chan, W., Conors, J. et al. (2003). The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. *Cancer Cell* 3 185–197.

van der Laan, M. & Bryan, J. (2001). Gene expression analysis with the parametric bootstrap. *Biostatistics* 2 445–461.

van der Laan, M. & Pollard, K. (2003). Hybrid clustering of gene expression data with visualization and the bootstrap. *Journal of Statistical Planning and Inference* 117 275–303.

van der Laan, M. & Robins, J. (2003). *Unified methods for censored longitudinal data and causality*. New York: Springer.

Appendix A

Efficient influence curve transformation for linear regression

The influence function for an estimator $\hat{\beta}_j$ of β_j in the regression model $Y = m(X_j|\beta_j) + \epsilon_j$ with $E(\epsilon_j|X_j) = 0$ is defined as the solution of

$$\sum_{i=1}^{n} h(\mathbf{X}(i,j), Y|\beta_j)) = 0$$

where $h(X_j, Y \mid \beta_j) = h(X_j)(Y - m(X_j|\beta_j))$ is the estimating function. The influence function is:

$$IC(X_j, Y) = -\left\{\frac{d}{d\beta_j} E(h(X_j, Y \mid \beta_j))\right\}^{-1} h(X_j, Y \mid \beta_j).$$

The efficient influence function uses the optimal estimating function h_{opt} (Bickel et al.(1993)Bickel, Klaassen, Ritov & Wellner). For example, in the case of linear regression, the optimal estimating equation is defined by $h_{opt}(X_j) = (1X_j)^\top / E(\epsilon_j^2(\beta) \mid X_j)$. Then,

$$0 = \sum_{i=1}^{n} \frac{(1\mathbf{X}(i,j))^\top}{E(\epsilon_j^2(\beta) \mid \mathbf{X}(i,j))} \epsilon(\beta)(i,j),$$

which corresponds with weighted least squares. For unweighted least squares $\sigma^2(X_j) = E(\epsilon_j^2(\beta) \mid X_j) = 1$, so the estimating function is simply $(1X_j)^\top \epsilon_j(\beta)$.

We now consider the linear regression model $E(Y \mid X_j) = \beta_{j0} + \beta_{j1} X_j$ with the intercept β_{j0} known in order to see that under a certain model we can view the influence curve for β_{j1} as a subject-specific slope. Under weak regularity conditions, we have that an efficient estimator $\hat{\beta}_{j1}$ of β_{j1} is asymptotically linear with an influence curve IC such that

$$\hat{\beta}_{j1} - \beta_{j1} = \frac{1}{n} \sum_{i=1}^{n} IC(\mathbf{X}(i,j), Y_i) + o_P(1/\sqrt{n}).$$

The efficient influence curve for the slope β_{j1} is

$$IC(X_j, Y) = \frac{1}{E(X_j^2/V(X_j))} \frac{X_j}{V(X_j)} \epsilon_j(\beta),$$

where $V(X_j) = \mathrm{VAR}(Y \mid X_j)$. In other words, in first order approximation we have:

$$\hat{\beta}_{j1} \approx \frac{1}{n} \sum_{i=1}^{n} \{\beta_{j1} + IC(\mathbf{X}(i,j), Y_i)\},$$

which shows that we might be able to view $\beta_{j1} + IC(\mathbf{X}(i,j), Y_i)$ as a subject-specific regression coefficient, whose average across subjects gives the overall regression coefficient.

To confirm this result, suppose $Y_i = \beta_{j0} + B_1(i,j)\mathbf{X}(i,j)$ (with no error) and β_{j1} is the overall regression coefficient. Then, $\epsilon(\beta)(i,j) = (B_1(i,j) - \beta_{j1})\mathbf{X}(i,j)$. So, the standardized residual transformation $W_j = \epsilon_j(\beta)/X_j$ is equal to $B_1(i,j) - \beta_{j1}$, which is the difference between the subject-specific slope and the overall slope. This provides some insight into the standardized residual transformation as an association profile, and provides a link between it and the influence curve for β_{j1}. Returning to the influence curve transformation, we have

$$\begin{aligned} V(\mathbf{X}(i,j)) &= E(\epsilon_j^2(\beta) \mid \mathbf{X}(i,j)) \\ &= \mathbf{X}(i,j)^2 E(B_1(i,j) - \beta_{j1})^2 \\ &\equiv \mathbf{X}(i,j)^2 \sigma_j^2(B). \end{aligned}$$

So $E(X_j^2/V(X_j)) = 1/\sigma_j^2(B)$ and $X_j/V(X_j) = 1/(X_j\sigma_j^2(B))$. In this case $IC(\mathbf{X}(i,j), Y_i \mid \beta_j) = (B_1(i,j) - \beta_{j1})$. Thus, $\beta_{j1} + IC(\mathbf{X}(i,j), Y_i \mid \beta_j) = B_1(i,j)$, the subject-specific slope in the model $Y_i = \beta_{j0} + B_1(i,j)\mathbf{X}(i,j)$ with no error. Note that this subject-specific contribution $B_1(i,j) = (Y_i - \beta_{j0})/\mathbf{X}(i,j)$ is independent of β_{j1}. So even when the sample size is low so that $\hat{\beta}_{j1}$ is a bad estimator of β_{j1} we will still obtain the exact subject-specific regression coefficient $B_1(i,j)$.

A similar calculation can be done for the influence curve in the model where both β_{j0} and β_{j1} are unknown.

Appendix B

Efficient influence function transformation for generalized linear regression

Suppose $E(Y \mid X_j) = m(X_j \mid \beta_j)$, where $\beta_j = (\beta_{j0}, \beta_{j1})$. Since β_{j0} is treated fixed in the following, for notational convenience, let $m(X_j \mid \beta_{j1})$ denote $m(X_j \mid \beta_j)$ and $m^1(X_j \mid \beta_{j1}) = \frac{d}{d\beta_{j1}} m(X_j \mid \beta_{j1})$. The efficient influence curve of β_{j1} in the model with β_{j0} known is given by:

$$IC(X_j, Y \mid \beta_{j1}) = \frac{1}{E(h_{opt}(X_j)m^1(X_j \mid \beta_{j1}))} h_{opt}(X_j)\epsilon_j(\beta_{j1}),$$

where $h_{opt}(X_j) \equiv \frac{m^1(X_j|\beta_{j1})}{E(\epsilon_j^2(\beta_{j1})|X_j)}$. Suppose $Y_i = m(\mathbf{X}(i,j) \mid B_1(i,j))$, where $B_1(i,j)$ is a random subject-specific coefficient whose variance we denote with $\sigma_j^2(B_1)$. Then

$$\begin{aligned} \epsilon_{ij}(\beta_{j1}) &= m(\mathbf{X}(i,j) \mid B_1(i,j)) - m(\mathbf{X}(i,j) \mid \beta_{j1}) \\ &= m^1(\mathbf{X}(i,j) \mid \beta_{j1})(B_1(i,j) - \beta_{j1}) + o_P(\mid B_1(i,j) - \beta_{j1} \mid). \end{aligned}$$

So, in first order, $E(\epsilon_j^2(\beta_{j1}) \mid X_j) = \sigma_j^2(B_1)\left\{m^1(X_j \mid \beta_{j1})\right\}^2$. This shows that $h_{opt}(X_j) \approx \frac{1}{\sigma_j^2(B_1)m^1(X_j|\beta_{j1})}$ and, since $E(h_{opt}(X_j)m^1(X_j \mid \beta_{j1})) = 1/\sigma_j^2(B_1)$, this proves $IC(\mathbf{X}(i,j), Y_i \mid \beta_{j1}) \approx B_1(i,j) - \beta_{j1}$. Thus, to a first order approximation, $\beta_{j1} + IC(\mathbf{X}(i,j), Y_i)$ can again be viewed as a subject-specific regression coefficient. Furthermore, for subject i the simple transformation $W_j = \epsilon_j(\beta)/m^1(X_j \mid \beta_j)$ is approximately equal to $B_1(i,j) - \beta_{j1}$, again suggesting an association profile that measures subject-specific deviations from the overall association β_{j1}.

Appendix C

Results for uncensored data

The proof of Theorem 1 for the residual transformation relies on convergence of $\hat{\mathbf{W}}$ to $\mathbf{W}$, which we prove first.

Lemma 1. *Let $W_j = Y - (\beta_{j0} + \beta_{j1}X_j)$, and suppose $|X_j| \leq M$ for a constant $M > 0$ and for $j = 1, \ldots, p$. If $\sup_j |\hat{\beta}_j - \beta_j| \xrightarrow[\frac{n}{\log p}\to\infty]{P} 0$, then*

$$\sup_j |\hat{W}_j - W_j| \xrightarrow[\frac{n}{\log p}\to\infty]{P} 0.$$

Proof.

$$\begin{aligned}
& \hat{W}_j - W_j \\
= \;& W_j(\hat{\beta}_j) - W_j(\beta_j) \\
= \;& Y - (\hat{\beta}_{j0} + \hat{\beta}_{j1}X_j) - \{Y - (\beta_{j0} + \beta_{j1}X_j)\} \\
= \;& (\beta_{j0} - \hat{\beta}_{j0}) + (\beta_{j1} - \hat{\beta}_{j1})X_j \\
\leq \;& \sup_j |\beta_{j0} - \hat{\beta}_{j0}| + \sup_j |X_j| * \sup_j |\beta_{j1} - \hat{\beta}_{j1}| \\
\leq \;& \sup_j |\beta_{j0} - \hat{\beta}_{j0}| + M * \sup_j |\beta_{j1} - \hat{\beta}_{j1}| \xrightarrow[\frac{n}{\log p}\to\infty]{P} 0.
\end{aligned}$$

So, $\sup_j |\hat{W}_j - W_j| \xrightarrow[\frac{n}{\log p}\to\infty]{P} 0$.

Proof of Theorem 1: Residual transformation

Let $W_j = Y - (\beta_{j0} + \beta_{j1}X_j)$, where Y is not censored. Suppose $|X_j| \leq M$ for a constant $M > 0$, $j = 1, \ldots, p$. If $\sup_j |\hat{\beta}_j - \beta_j| \xrightarrow[\frac{n}{\log p}\to\infty]{P} 0$, then $\sup_{j,j'} |\hat{\mathbf{D}}(j,j') - \mathbf{D}(j,j')| \xrightarrow[\frac{n}{\log p}\to\infty]{P} 0$.

Proof. The theorem follow directly from Lemma 1, since $\mathbf{D}$ is a deterministic function of $\mathbf{W}$ and $\hat{\mathbf{D}}$ is a deterministic function of $\hat{\mathbf{W}}$.

Proof of Theorem 3:

Let $W_j = Y - (\beta_{j0} + \beta_{j1} X_j)$, where Y is not censored. Consider the estimator $\bar{\mathbf{D}} = \mathbf{D}(\hat{\mu}, \hat{\Sigma})$ of $\mathbf{D} = \mathbf{D}(\mu, \Sigma)$, defined above. Suppose $|X_j| \leq M$, $|Y| \leq M$, $|W_j| \leq M$, $|\beta_j| \leq M$, and $|\hat{\beta}_j| \leq M$ for a constant $M > 0$, $j = 1, \ldots, p$. If the variance of X_j is bounded away from zero uniformly in j, then $\sup_{j,j'} |\bar{\mathbf{D}}(jj') - \mathbf{D}(jj')| \xrightarrow[\frac{n}{\log p} \to \infty]{P} 0$.

Proof. Recall that $\mu_j = E(W_j)$ and $\sigma_{jj'} = E(W_j W_{j'})$. Consider two estimators for each of these moments of the distribution of W_j: (i) the empirical estimate based on the true transformation (*i.e.* if the regression coefficients are known) and (ii) the empirical estimate based on an estimated transformation (*i.e.* if the regression coefficients are estimated). Denote these estimators by

$$\begin{aligned} \tilde{\mu}_j &= \frac{1}{n} \sum_{i=1}^{n} \mathbf{W}(i,j) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{W}(\beta)(i,j) \\ \tilde{\sigma}_{jj'} &= \frac{1}{n} \sum_{i=1}^{n} \mathbf{W}(i,j)\mathbf{W}(i,j') = \frac{1}{n} \sum_{i=1}^{n} \mathbf{W}(\beta)(i,j)\mathbf{W}(\beta)(i,j') \end{aligned}$$

for the true transformation $\mathbf{W} = \mathbf{W}(\beta)$ and

$$\begin{aligned} \hat{\mu}_j &= \frac{1}{n} \sum_{i=1}^{n} \hat{\mathbf{W}}(i,j) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{W}(\hat{\beta})(i,j) \\ \hat{\sigma}_{jj'} &= \frac{1}{n} \sum_{i=1}^{n} \hat{\mathbf{W}}(i,j)\hat{\mathbf{W}}(i,j') = \frac{1}{n} \sum_{i=1}^{n} \mathbf{W}(\hat{\beta})(i,j)\mathbf{W}(\hat{\beta})(i,j') \end{aligned}$$

for the estimated transformation $\hat{\mathbf{W}} = \mathbf{W}(\hat{\beta})$. We need to show that under the conditions of the theorem:

1. $\sup_j |\hat{\mu}_j - \mu_j| \xrightarrow[\frac{n}{\log p} \to \infty]{P} 0$,
2. $\sup_{j,j'} |\hat{\sigma}_{jj'} - \sigma_{jj'}| \xrightarrow[\frac{n}{\log p} \to \infty]{P} 0$.

Then, the result of the theorem follows, since $\mathbf{D}$ and $\bar{\mathbf{D}}$ are deterministic functions of (μ, Σ) and $(\hat{\mu}, \hat{\Sigma})$, respectively.

Proof of 1:

$$
\begin{aligned}
& \hat{\mu}_j - \mu_j \\
= \;& (\hat{\mu}_j - \tilde{\mu}_j) + (\tilde{\mu}_j - \mu_j) \\
= \;& \frac{1}{n}\sum_{i=1}^{n}\Big\{\mathbf{W}(\hat{\beta})(i,j) - \mathbf{W}(\beta)(i,j)\Big\} + (\tilde{\mu}_j - \mu_j) \\
= \;& \frac{1}{n}\sum_{i=1}^{n}\Big\{(Y_i - \hat{\beta}_{j0} - \hat{\beta}_{j1}\mathbf{X}(i,j)) - (Y_i - \beta_{j0} - \beta_{j1}\mathbf{X}(i,j))\Big\} + (\tilde{\mu}_j - \mu_j) \\
= \;& \frac{1}{n}\sum_{i=1}^{n}\Big\{(\beta_{j0} - \hat{\beta}_{j0}) + \mathbf{X}(i,j)(\beta_{j1} - \hat{\beta}_{j1})\Big\} + (\tilde{\mu}_j - \mu_j) \\
\leq \;& \sup_j \bar{X}_j * \left\{\sup_j |\beta_{j0} - \hat{\beta}_{j0}| + \sup_j |\beta_{j1} - \hat{\beta}_{j1}|\right\} + \sup_j |\tilde{\mu}_j - \mu_j| \\
\leq \;& M * \left\{\sup_j |\beta_{j0} - \hat{\beta}_{j0}| + \sup_j |\beta_{j1} - \hat{\beta}_{j1}|\right\} + \sup_j |\tilde{\mu}_j - \mu_j| \xrightarrow[\frac{n}{\log p}\to\infty]{P} 0.
\end{aligned}
$$

For a proof of the convergence of $\sup_j |\beta_{j0} - \hat{\beta}_{j0}|$ and $\sup_j |\beta_{j1} - \hat{\beta}_{j1}|$ we refer the reader to Bryan(2001)) (p.52-53). The convergence of $\sup_j |\tilde{\mu}_j - \mu_j|$ follows a similar Berstein's inequality argument. Since $\hat{\mu}_j - \mu_j$ converges to zero, so does $\sup_j |\hat{\mu}_j - \mu_j|$, completing the proof.

Proof of 2:

$$
\begin{aligned}
& \hat{\sigma}_{jj'} - \sigma_{jj'} \\
= \; & (\hat{\sigma}_{jj'} - \tilde{\sigma}_{jj'}) + (\tilde{\sigma}_{jj'} - \sigma_{jj'}) \\
= \; & \frac{1}{n}\sum_{i=1}^{n}\Big\{\mathbf{W}(\hat{\beta})(i,j)\mathbf{W}(\hat{\beta})(i,j') - \mathbf{W}(\beta)(i,j)\mathbf{W}(\beta)(i,j')\Big\} + (\tilde{\sigma}_{jj'} - \sigma_{jj'}) \\
= \; & \frac{1}{n}\sum_{i=1}^{n}\Big\{(Y_i - \hat{\beta}_{j0} - \hat{\beta}_{j1}\mathbf{X}(i,j))(Y_i - \hat{\beta}_{j'0} - \hat{\beta}_{j'1}\mathbf{X}(i,j'))\Big\} \\
& -\frac{1}{n}\sum_{i=1}^{n}\left\{(Y_i - \beta_{j0} - \beta_{j1}\mathbf{X}(i,j))(Y_i - \beta_{j'0} - \beta_{j'1}\mathbf{X}(i,j'))\right\} \\
& +(\tilde{\sigma}_{jj'} - \sigma_{jj'}) \\
= \; & \frac{1}{n}\sum_{i=1}^{n} Y_i\Big\{(\beta_{j0} - \hat{\beta}_{j0}) + (\beta_{j'0} - \hat{\beta}_{j'0}) + \mathbf{X}(i,j)(\beta_{j1} - \hat{\beta}_{j1}) + \mathbf{X}(i,j')(\beta_{j'1} - \hat{\beta}_{j'1})\Big\} \\
& -\frac{1}{n}\sum_{i=1}^{n}\Big\{\mathbf{X}(i,j')(\beta_{j0}\beta_{j'1} - \hat{\beta}_{j0}\hat{\beta}_{j'1}) + \mathbf{X}(i,j)(\beta_{j'0}\beta_{j1} - \hat{\beta}_{j'0}\hat{\beta}_{j1}) + \mathbf{X}(i,j)\mathbf{X}(i,j')(\beta_{j1}\beta_{j'1} - \hat{\beta}_{j1}\hat{\beta}_{j'1})\Big\} \\
& -(\beta_{j0}\beta_{j'0} - \hat{\beta}_{j0}\hat{\beta}_{j'0}) + (\tilde{\sigma}_{jj'} - \sigma_{jj'}) \\
\leq \; & 2M * \sup_j |\hat{\beta}_{j0} - \beta_{j0}| + 2M^2 * \sup_j |\hat{\beta}_{j1} - \beta_{j1}| + (\hat{\beta}_{j0} - \hat{\beta}_{j1}\bar{X}_j)(\hat{\beta}_{j'0} - \hat{\beta}_{j'1}\bar{X}_{j'}) \\
& -(\beta_{j0} - \beta_{j1}\bar{X}_j)(\beta_{j'0} - \beta_{j'1}\bar{X}_{j'}) + (\tilde{\sigma}_{jj'} - \sigma_{jj'}) \\
\leq \; & 2M * \sup_j |\hat{\beta}_{j0} - \beta_{j0}| + 2M^2 * \sup_j |\hat{\beta}_{j1} - \beta_{j1}| + 2M^2 + 4M^3 + 2M^4 + (\tilde{\sigma}_{jj'} - \sigma_{jj'}) \\
& \xrightarrow[\frac{n}{\log p}\to\infty]{P} 0.
\end{aligned}
$$

So, $\sup_{j,j'} |\hat{\sigma}_{jj'} - \sigma_{jj'}| \xrightarrow[\frac{n}{\log p}\to\infty]{P} 0$. Again, we repeatedly use the Bernstein's inequality result from Bryan(2001))(p.52-53). The requirement that $var(X_j)$ be bounded away from zero is needed for distances, such as correlation distance, where one must divide by $\sigma_{jj'}$.

Appendix D

Results for censored data

Proof of Theorem 4:

Let $W_j = Y - (\beta_{j0} + \beta_{j1}X_j)$, where Y is right-censored. Consider the estimator $\bar{\mathbf{D}} = \mathbf{D}(\hat{\mu}, \hat{\Sigma})$ of $\mathbf{D} = \mathbf{D}(\mu, \Sigma)$, defined above. Suppose $|X_j| \leq M$, $|Y| \leq M$, $|W_j| \leq M$, $|\beta_j| \leq M$, and $|\hat{\beta}_j| \leq M$

for a constant $M > 0$, $j = 1, \ldots, p$. Also suppose the variance of X_j is bounded away from zero uniformly in j. Assume that (i) $C \perp Y|X$, (ii) $\sup |\bar{G}_n(Y|\mathbf{X}, Y) - \bar{G}(Y|\mathbf{X}, Y)| \xrightarrow{P} 0$ where the supremum is over the support of the distribution of $(\mathbf{X}, Y)$, and (iii) $\bar{G}(Y|\mathbf{X}, Y) > \delta > 0$ for a.e. $(\mathbf{X}, Y)$, where $\bar{G}(t|\mathbf{X}, Y) = Pr(C > t|\mathbf{X}, Y)$. Then, $\sup_{j,j'} |\bar{\mathbf{D}}(j, j') - \mathbf{D}(j, j')| \xrightarrow[\frac{n}{\log p} \to \infty]{P} 0$.

Proof. We need the same two convergence conditions as in the proof of Theorem 3 where (μ, Σ) are now estimated by the IPCW estimators of Equations 2 and 3 and W_j is not observed for all subjects. In otherwords, we need to show:

1. $\sup_j |\hat{\mu}_j - \mu_j| = \sup_j |\frac{1}{n} \sum_{i=1}^n \frac{\Delta_i \mathbf{W}(\hat{\beta})(i,j)}{\bar{G}_n(Y_i|\mathbf{X},Y)} - \mu_j| \xrightarrow[\frac{n}{\log p} \to \infty]{P} 0$,
2. $\sup_{jj'} |\hat{\sigma}_{jj'} - \sigma_{jj'}| = \sup_{j,j'} |\frac{1}{n} \sum_{i=1}^n \frac{\Delta_i \mathbf{W}(\hat{\beta})(i,j)\mathbf{W}(\hat{\beta})(i,j')}{\bar{G}_n(Y_i|\mathbf{X},Y)} - \sigma_{jj'}|$
   $\xrightarrow[\frac{n}{\log p} \to \infty]{P} 0$.

Then, the theorem follows, since $\mathbf{D}$ and $\bar{\mathbf{D}}$ are deterministic functions of (μ, Σ) and $(\hat{\mu}, \hat{\Sigma})$, respectively.

Proof of 1:

$$
\begin{aligned}
& \hat{\mu}_j - \mu_j \\
= \; & \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\Delta_i}{\bar{G}_n(Y_i|\mathbf{X},Y)}(\mathbf{W}(\hat{\beta})(i,j)-\mathbf{W}(\beta)(i,j))\right\}+\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\Delta_i\mathbf{W}(\beta)(i,j)}{\bar{G}_n(Y_i|\mathbf{X},Y)}-\mu_j\right\} \\
= \; & \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\Delta_i}{\bar{G}_n(Y_i|\mathbf{X},Y)}(\hat{\beta}_{j0}-\beta_{j0}+\mathbf{X}(i,j)(\hat{\beta}_{j1}-\beta_{j1}))\right\}+\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\Delta_i\mathbf{W}(\beta)(i,j)}{\bar{G}_n(Y_i|\mathbf{X},Y)}-\mu_j\right\} \\
= \; & (\hat{\beta}_{j0}-\beta_{j0})*\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\bar{G}_n(Y_i|\mathbf{X},Y)}+(\hat{\beta}_{j1}-\beta_{j1})*\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i\mathbf{X}(i,j)}{\bar{G}_n(Y_i|\mathbf{X},Y)} \\
& +\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\Delta_i\mathbf{W}(\beta)(i,j)}{\bar{G}_n(Y_i|\mathbf{X},Y)}-\mu_j\right\} \\
= \; & (\hat{\beta}_{j0}-\beta_{j0})*\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\bar{G}_n(Y_i|\mathbf{X},Y)}+(\hat{\beta}_{j1}-\beta_{j1})*\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i\mathbf{X}(i,j)}{\bar{G}_n(Y_i|\mathbf{X},Y)} \\
& +\frac{1}{n}\sum_{i=1}^{n}\left\{\Delta_i\mathbf{W}(\beta)(i,j)\left(\frac{1}{\bar{G}_n(Y_i|\mathbf{X},Y)}-\frac{1}{\bar{G}(Y_i|\mathbf{X},Y)}\right)\right\}+\sum_{i=1}^{n}\left\{\frac{\Delta_i\mathbf{W}(\beta)(i,j)}{\bar{G}(Y_i|\mathbf{X},Y)}-\mu_j\right\} \\
= \; & (\hat{\beta}_{j0}-\beta_{j0})*\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\bar{G}_n(Y_i|\mathbf{X},Y)}+(\hat{\beta}_{j1}-\beta_{j1})*\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i\mathbf{X}(i,j)}{\bar{G}_n(Y_i|\mathbf{X},Y)} \\
& +\frac{1}{n}\sum_{i=1}^{n}\left\{\Delta_i\mathbf{W}(\beta)(i,j)\frac{\bar{G}(Y_i|\mathbf{X},Y)-\bar{G}_n(Y_i|\mathbf{X},Y)}{\bar{G}_n(Y_i|\mathbf{X},Y)\bar{G}(Y_i|\mathbf{X},Y)}\right\}+(\tilde{\mu}_j-\mu_j)
\end{aligned}
$$

By our assumptions

$$Pr\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i\mathbf{X}(i,j)}{\bar{G}_n(Y_i|\mathbf{X},Y)}<C\right)\to 1$$

and

$$Pr\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\bar{G}_n(Y_i|\mathbf{X},Y)}<C\right)\to 1$$

for some $C<\infty$. So, we have:

$$
\begin{aligned}
& \hat{\mu}_j - \mu_j \\
\leq \; & C*(\sup_j|\beta_{j0}-\hat{\beta}_{j0}|+\sup_j|\beta_{j1}-\hat{\beta}_{j1}|)+\sup_j\left|\frac{\bar{G}_n(Y|\mathbf{X},Y)-\bar{G}(Y|\mathbf{X},Y)}{\bar{G}_n(Y_i|\mathbf{X},Y)\bar{G}(Y_i|\mathbf{X},Y)}\right|*\frac{1}{n}\sum_{i=1}^{n}|\Delta_i\mathbf{W}(\beta)(i,j)| \\
& +\sup_j|\tilde{\mu}_j-\mu_j| \\
\leq \; & C*(\sup_j|\beta_{j0}-\hat{\beta}_{j0}|+\sup_j|\beta_{j1}-\hat{\beta}_{j1}|)+\sup_j|\mathbf{W}(\beta)(i,j)|*O_p(1/\sqrt{n})+\sup_j|\tilde{\mu}_j-\mu_j| \\
\leq \; & C*(\sup_j|\beta_{j0}-\hat{\beta}_{j0}|+\sup_j|\beta_{j1}-\hat{\beta}_{j1}|)+M*O_p(1/\sqrt{n})+\sup_j|\tilde{\mu}_j-\mu_j| \xrightarrow[\frac{n}{\log p}\to\infty]{P} 0.
\end{aligned}
$$

Again, the convergence of $\sup_j |\beta_{j0} - \hat{\beta}_{j0}|$, $\sup_j |\beta_{j1} - \hat{\beta}_{j1}|$ and $\sup_j |\tilde{\mu}_j - \mu_j|$ follow from Bernstein's inequality.

We omit the proof of part 2 (second moments). The argument combines the ideas from the proof of part 1 of this theorem with those of part 2 of Theorem 3. Again, Bernstein's inequality is used repeatedly.

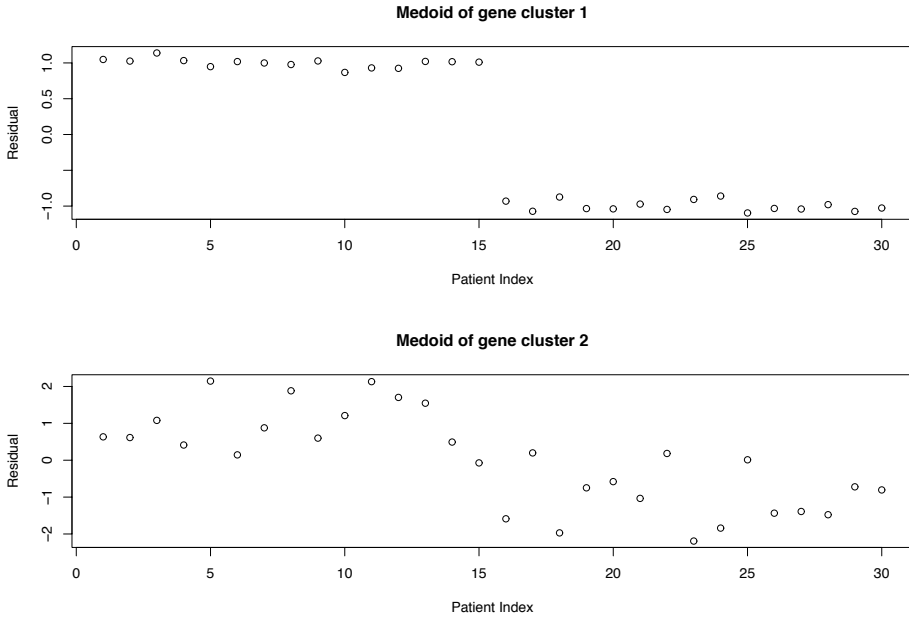

Figure 1. Plots of the residuals in simulated data. Residuals for each patient are computed from the regression of two different genes on $\log T$, one depicted in each panel. The top panel is the medoid gene for the small cluster identified in the analysis of residual transformed data (containing the causal gene). The bottom panel is the medoid gene for the larger cluster from that analysis. The patients separate clearly into two groups in the top panel, but not the bottom panel.

Figure 2. Distance matrices from simulated censored data. Euclidean gene $\times$ gene distance matrices, with genes ordered by clusters. Each pairwise distance is represented by a color on the red-white scale, with bright red corresponding to the smallest distance. Both panels are from the analysis of the residual transformed data with the PAM clustering algorithm. There are $k = 2$ clusters in each panel. The left panel is the IPCW estimated matrix with $\approx 20\%$ censoring. The right panel is with $\approx 30\%$ censoring. The clusters are less distinct with more censoring.

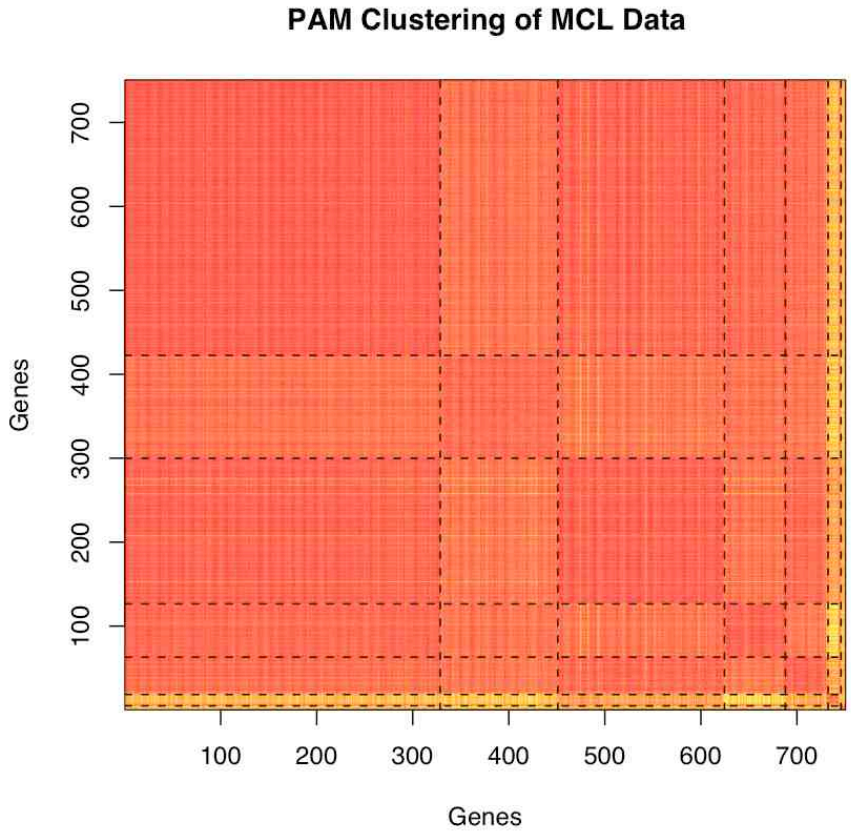

Figure 3. Distance matrix from MCL censored data. Euclidean gene × gene distance matrices, with genes ordered by clusters. Distance matrix is based on the residual transformation and is estimated with IPCW estimators, due to 30.4% censoring. There are $k = 7$ clusters, whose boundaries are marked with dashed lines. The color scheme is the same as in Figure 2.

Transformation	**Inspiration**	**Notes**
β_{j1}	Regression coefficient	Scalar, not n-vector
ϵ_j	Residuals	
$\frac{1}{E(X_j^2/V(X_j))}\frac{X_j}{V(X_j)}\epsilon_j$	Influence curve	$V(X_j) = \mathrm{VAR}(Y\|X_j)$
$\epsilon_j/\tilde{X}_j$	Standardized residuals	$\tilde{X}_j$ is X_j, possibly bounded away from 0

Table 1

Examples of transformations $W_j = W_j(\mathbf{X}, Y)$ for the linear regression model $Y = \beta_{j0} + \beta_{j1}X_j + \epsilon_j$.