Statistica Sinica **12**(2002), 263-272

MODELING THE INFLUENCE OF DISEASE PROCESSES ON GENE EXPRESSION

Kerby Shedden

University of Michigan

Abstract: We introduce a new approach for modeling the influence of a disease process on gene or protein expression patterns. Our emphasis is on the simultaneous, multivariate characterization of expression alterations across large numbers of genes, rather than on the construction of normal/affected differential expression profiles one gene at a time. The key idea is to reconstruct the expression profile for a latent sample of normal control cells corresponding to each disease sample, and then use the displacements between the disease samples and their reconstructed controls to uncover a low-dimensional range of alternative disease progression pathways. The method is easy to implement, and is expected to be widely-applicable to genomic and proteomic studies using a broad range of large-scale assay technologies. We demonstrate the method by applying it to gene expression studies of colon cancer and breast cancer.

Key words and phrases: Cancer, differential expression, dimension reduction, gene expression, tumor progression.

1. Introduction

Most diseases exhibit a range of clinical manifestations. Underlying this clinical heterogeneity is expected to be a correspondingly heterogeneous range of behaviors at the level of gene and protein expression. Arriving at an understanding of the extent and boundaries of this variability has become a major research goal as new technologies for carrying out large-scale gene and protein expression assays, such as cDNA microarrays and 2D gel electrophoresis, have come into widespread use. In this paper, we address the problem of statistically characterizing the diversity of gene expression alterations that are associated with the progression of a disease. Our approach borrows from dimension reduction methods for high dimensional data analysis, and differs from most other approaches to this problem in that it aims to identify the multivariate structure of expression alterations simultaneously across the entire range of measured genes. Specifically, if $\{T_j\}$ and $\{N_j\}$ are independent sets of disease and normal samples, we aim to look beyond the average difference $\bar{T} - \bar{N}$, so as to uncover the range of discrepancies $T_j - \tilde{T}_j$ that can occur between a disease sample T_j, and a hypothetical paired sample $\tilde{T}_j$ of normal control cells.

To fix ideas, we focus on modeling tumor progression in cancer. Suppose that $Y(t) \in \mathcal{R}^n$ contains the expression measurements for n genes in a sample of tumor cells at time t, where $t = 0$ corresponds to tumor initiation. Let $\nabla_t Y$ denote the vector of first derivatives of $Y(t)$ with respect to time. If we require that for all t, $\nabla_t Y \in \langle \nu_1, \ldots, \nu_p \rangle$ where $\nu_j \in \mathcal{R}^n$, then $Y(t) - Y(0)$ will lie in $\langle \nu_1, \ldots, \nu_p \rangle$ for all t. In principle, we can take $p = n$, and the model imposes no constraints on the change in gene expression over time. In practice, a desire to achieve parsimony motivates us to consider much smaller values for p, leading to the following model for disease progression:

$$Y(t) = Y(0) + \sum_{k=1}^{p} \lambda_k(t)\nu_k + \epsilon(t), \tag{1}$$

where $\epsilon(t)$ is an error component with mean zero given $\lambda_1(t), \ldots, \lambda_p(t)$.

In the context of (1), we can view the ν_k as representing alternative pathways of disease progression. Since $\nu_k(i)$ represents the rate at which gene i is modified while following pathway ν_k, we refer to the ν_k as the *slope vectors*, and the subspace $\langle \nu_1, \ldots, \nu_d \rangle$ as the *slope subspace*. We take the viewpoint that the ν_k are shared across all samples that belong to a broadly-defined disease class, while the $\lambda_k(t)$ represent the molecular progressions of individual tumors, and hence will be specific to each sample. Note that $\lambda_k(t) = 0$ corresponds to gene expression in the normal precursor cell, while larger values of $\lambda_k(t)$ represent increasing levels of genetic transformation and deranged expression behavior.

To better understand the biological basis for disease progression, we briefly describe the process of tumor development in cancer (Lengauer, Kinzler and Vogelstein (1998), Vogelstein and Kinzler (1993)). It is generally believed that most tumors derive from a single normal ancestor cell that has experienced a genetic transformation known as tumor initiation. While a single mutation cannot give rise to a malignant tumor, it can set the stage for further mutations. For example, the mutation may inhibit the DNA proofreading and repair mechanisms that maintain the status of each somatic cell as a clone of all other cells in the organism. In the next stage, a genetically heterogeneous population of tumor cells forms, and natural selection begins to exert a selective pressure on the population. This pressure favors cancer-like cells having traits such as rapid division, invasive capabilities and immune resistance, while acting against more mildly transformed cells that terminally differentiate into the functional role of the untransformed progenitor. Therefore a primary aspect of tumor development is the gradual loss of differentiation, which can be measured through discrepancies in gene expression between tumor cells and normal control cells of the type that is thought to have given rise to the tumor. This normal expression behavior constitutes the natural baseline for assessing expression discrepancies that result

from cancer. Since the baseline expression behavior will vary from individual to individual, we should treat it as a tumor-specific control.

If paired disease and normal samples are available (for example if a tissue sample has been taken from a healthy region of the affected organ, or from the contralateral organ), then the estimation of model (1) is not difficult. Suppose Y_{j0} and Y_{j1} are paired normal and disease samples, and let X denote the matrix whose columns are the differences $Y_{j1} - Y_{j0}$, where j ranges over the observed sample pairs. Next apply the singular value decomposition (SVD) to X, giving $X = USV'$, where U and V are orthogonal, and S is diagonal with decreasing diagonal entries. The columns of U corresponding to large diagonal elements of S estimate the ν_k, and the sample-specific coefficients $\lambda_{kj}(t)$ are estimated by the values $S_{kk}V_{jk}$. We note that most commonly, each disease sample is observed only a single time, so the effect of time is completely confounded with subject-specific characteristics of the sample. Therefore we are only able to estimate the λ values at the tumor level, giving estimates $(\hat{\lambda}_{1j}, \ldots, \hat{\lambda}_{pj})$ that represent the progression of disease sample j at the time that the RNA or protein for disease sample j was obtained.

In general paired samples will not be available, rather expression measurements $\{N_j\}$ and $\{T_j\}$ corresponding to independent samples of normal and tumor cells will be obtained. We posit that corresponding to each tumor sample T_j is a latent normal expression profile $\tilde{T}_j$, representing a sample-specific normal precursor, or control, to tumor sample j (this would correspond to the expression levels $Y(0)$ in the continuous measurement setting). The $\tilde{T}_j$ will vary from subject to subject due to individual-specific health factors. We model the expected values of the tumor samples as lying in an affine subspace of $\mathcal{R}^n$ given by $\mu_N + \langle \mathcal{B}_N \rangle$, where $\mathcal{B}_N$ is an $n \times d$ matrix, and $\langle \mathcal{B}_N \rangle$ represents the column-space of $\mathcal{B}_N$. Since our only information about the normal expression behavior comes from the set of independent normal samples $\{N_j\}$, we estimate μ_N as the average normal expression measurement $\bar{N}$, and $\mathcal{B}_N$ as the matrix whose columns are the d dominant left singular vectors of the matrix $\mathcal{N}_C = [N_1 - \bar{N}, N_2 - \bar{N}, \cdots]$. The matrix $\mathcal{B}_N$ obtained in this way is identical to the matrix that would be obtained by performing a principal components analysis (PCA) on the covariance matrix of the normal samples, but is much easier to compute since the covariance matrix is substantially larger than $\mathcal{N}_C$. This leads to the following model for $\tilde{T}_j$, where β_N^j is the unobservable coefficient vector that determines the sample-specific characteristics of normal sample j:

$$\tilde{T}_j = \mu_N + \mathcal{B}_N \beta_N^j + \tau_j, \tag{2}$$

where $E(\tau_j | \beta_N^j) = 0$.

The observed gene expression for disease sample j is modeled as the sum of the unobserved normal precursor $\tilde{T}_j$ and the influence of disease progression,

which we represent as a linear combination of a low dimensional set of alternative disease progression pathways. This leads to the following model:

$$T_j = \tilde{T}_j + \sum_{k=1}^{p} \lambda_{kj}\nu_k + \epsilon_j, \tag{3}$$

where $E(\epsilon_j|\lambda_{1j}, \ldots, \lambda_{pj}) = 0$.

We can view the model given by (3) as capturing a tradeoff between two different sources of disease heterogeneity. At one extreme, we have diseases that arise from normal cells exhibiting a very diverse range of expression behavior, but that are highly specific in the way that the expression behavior is altered by the disease process. This would require a large value of d (a high-dimensional $\mathcal{B}_N$) and a small value for p (a small range of alternative development pathways). At the other extreme we have diseases that arise from a normal cell type whose gene expression is very tightly controlled, but that are variable in the types of expression modifications that can occur, leading to small values of d and larger values for p.

Estimation of model (3) can be carried out using least squares. In order for the coefficients to be identified, we require that $\mathcal{B}_N'\nu_k = 0$ for $k = 1, \ldots, p$. The loss function

$$\mathcal{L}(\{\beta_N^j\}, \{\lambda_{kj}\}, \{\nu_k\}) = \sum_j \|T_j - \mu_N - \mathcal{B}_N\beta_N^j - \sum_{k=1}^{p} \lambda_{kj}\nu_k\|^2 \tag{4}$$

has a unique global minimum that is obtained by the following two-stage procedure. First, regress $T_j - \mu_N$ on $\mathcal{B}_N$ to give β_N^j. Next perform a singular value decomposition on the matrix $\mathbf{R}$ whose columns are the residuals $R_j = T_j - \mathcal{B}_N\beta_N^j$, giving $\mathbf{R} = USV'$. The slope vectors ν_k are estimated as the first p columns of U, and the coefficients are estimated as $\lambda_{kj} = S_{kk}V_{jk}$. We note that we usually standardize each gene to have unit variance before fitting the model, so that all genes have equal influence on the fit. We also note that the differences $T_j - \tilde{T}_j$, and hence the ν and λ estimates, are invariant under location transforms such as mean-centering or median-centering of the genes.

Model (3) is not uniquely parameterized, so we cannot identify specific slope vectors $\nu_1, \ldots, \nu_p$ without imposing further constraints. Moreover, if n is smaller than p, a p-dimensional subspace of slope vectors cannot exist, so the ν_k will necessarily be linearly dependent. In practice, we can only identify a slope subspace that is a subspace of $\langle T_j - \tilde{T}_j; j = 1, \ldots, m_2\rangle$, which cannot have dimension greater than $\min(n, m_2)$. If the T_j and N_j are mutually linearly independent, then we can choose $1 \le p \le \min(n, m_2)$ and the minimization of (4) will define $\langle\nu_1, \ldots, \nu_p\rangle$ uniquely. The slope vectors themselves are taken for convenience to be those arising from the SVD. We note that, at the present time, the number

of genes that are measured usually exceeds the number of samples that can be obtained by an order of magnitude or more, so the minimizer of (4) is almost certain to be unique.

To summarize, we list the steps of our proposed procedure explicitly.

1. Compute the normal mean $\bar{N}$ and perform a PCA on the normal samples, producing a basis $\mathcal{B}_N$ for the variation in normal samples.
2. Regress $T_j - \bar{N}$ on $\mathcal{B}_N$ for each disease sample j, producing a reconstruction $\tilde{T}_j = \bar{N} + \mathcal{B}_N \beta_N^j$ of gene expression in the normal precursor to disease sample j.
3. Compute the singular value decomposition of the residual matrix $USV' = [R_1, R_2, \ldots]$, where $R_j = T_j - \tilde{T}_j$. Estimate the dimension p of the slope subspace using the values in S. Then estimate the slope vector ν_k as the k^{th} column of U, and the loading coefficients as $\lambda_{kj} = S_{kk} V_{jk}$.

2. Example: Colon Cancer

Our first example will use the colon carcinoma data reported by Notterman, Alon, Sierk and Levine (2001). In this study, paired normal and tumor tissue samples were obtained from 18 patients, and expression levels for 7464 genes were measured using oligonucleotide microarrays. Since there is a known pairing between tumor and normal samples, this data set will be useful for validating that our method performs in a manner consistent with its underlying motivation. Specifically, we will demonstrate that the slope vectors ν_k and the sample-specific coefficients λ_{kj} determined by our proposed method are roughly similar to the values obtained by directly fitting model (3) upon substituting the observed paired normal sample for $\tilde{T}_j$.

Figure 1 shows the estimated slopes for all 7464 genes, with the slope estimates for ν_1 shown in the left panel, those for ν_2 shown in the right panel. In both cases, the estimate derived from the known pairing provides the y-coordinate, while the estimate derived by reconstructing the control member of the pair using our proposed procedure provides the x-coordinate. Throughout this section, when using our reconstruction procedure we use $d = 3$ dimensions for determining the range of possible values for the normal reconstruction. The slope vector ν_1 is very similar in the two methods. We find that ν_1 is generally very similar to the displacement vector between the tumor centroid and the normal centroid, so this finding should not be unexpected, as the displacement vector is independent of the way the samples are paired. The slope vector ν_2 presents a much stronger test of our method. The correlation coefficient between the slopes in ν_2 obtained with the known pairing and the corresponding slopes obtained using the reconstructed pairing is 0.63 which, while not perfect, gives reasonable agreement for most genes.

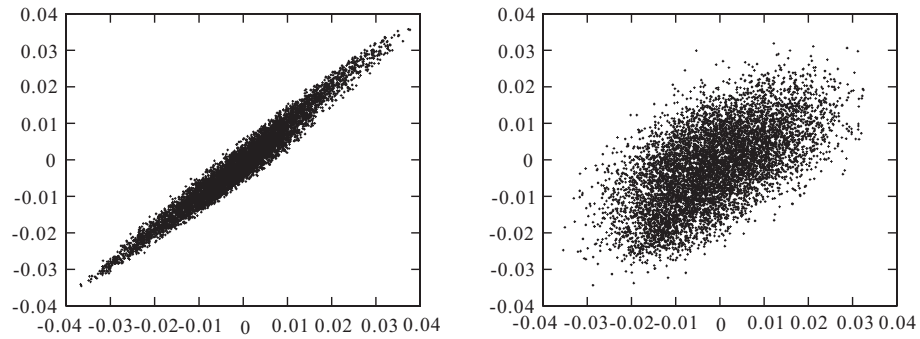

Figure 1. Component-wise scatterplots of the estimated slopes $\nu_k(i)$, $i = 1, \ldots, 7464$, plotting the estimate based on the known pairing for the colon tumor data (y-axis) against the estimate produced using our proposed reconstruction procedure (x-axis). The left panel shows the values for ν_1, and the right panel shows the values for ν_2.

The left panel of Figure 2 shows the λ_{1j} estimates produced using the known pairing plotted against the λ_{1j} estimates produced using our procedure to reconstruct the pairing. These points are shown using + symbols. On the same axes, the analogous plot for the λ_{2j} estimates is shown using $\times$ symbols. Again, while the agreement is not perfect, the basic relationships are preserved. The correlation between the two sets of λ_{1j} estimates is 0.61, while the analogous correlation for the λ_{2j} estimates is 0.72.

In the right panel of Figure 2, we show the scatterplot of $(\lambda_{1j}, \lambda_{2j})$, which would be a primary object of interest in practice. The tumor samples are seen to fall into two well-separated groups that are determined by the sign of λ_{2j}. The group of samples in the upper half-plane of the scatterplot will tend to have overexpression of genes with positive slopes in ν_2 and underexpression of genes with negative slopes in ν_2. The opposite tendency will hold for the samples in the lower half-plane of the scatterplot. For example, the greatest positive slope in ν_2 is gene R41349, while the most negative slope is for the APC gene. Thus samples in the upper half-plane will tend to underexpress APC and overexpress R41349, while the samples in the lower half-plane will tend to overexpress APC and underexpress R41349. We emphasize, however, that the slopes in component ν_2 are based on all 7464 genes, with several hundred genes having slopes that are substantially different from zero. Thus the values of λ_{2j} may not be tightly associated with any single gene, rather they reflect a consistent pattern of expression across a large group of genes.

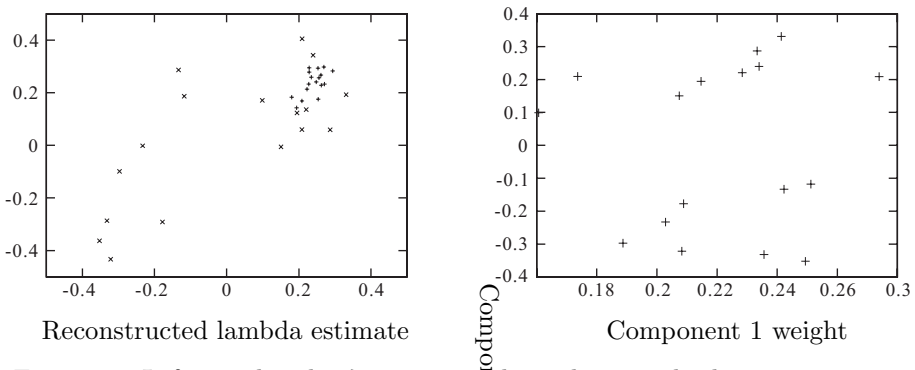

Figure 2. Left panel: The λ estimates derived using the known pairings plotted against the corresponding λ estimates derived using our proposed reconstruction procedure. The + symbols correspond to the λ_{1j} estimates, while the $\times$ symbols correspond to the λ_{2j} estimates. Right panel: Scatterplot of the $(\lambda_{1j}, \lambda_{2j})$ estimates produced by our procedure.

3. Example: Breast Cancer

Our second example will use the data from the breast tumor study reported in Perou et al. (2000). In this study, cDNA microarrays were used to assay the expression of 9216 genes in 3 normal breast tissue samples N_1, N_2, N_3 and 59 breast tumor samples $T_1, \ldots, T_{59}$. According to the authors, this study was intentionally planned to show the diversity of gene transcription in human breast tumors. As might be expected, there is a substantial amount of coordinated expression behavior that is not captured in the primary tumor/normal mean displacement vector. Our method uncovers some of this coordinated behavior, enabling us to link gene expression in the breast tumors to a clinical variable that would otherwise be difficult to characterize.

We began by applying the method directly as described in Section 1. Around 3% of the spots were missing, so we filled in these values with the average of all observed values for the same gene. Since there are only three observed normal samples, we used the maximal value of $d = 2$ for determining the range of possible values for the normal reconstruction. The left panel of Figure 3 shows a scatterplot of the pairs $(\lambda_{1j}, \lambda_{2j})$. As was the case with the tumor samples, the λ_{1j} are all positive. This indicates that the displacement vectors between tumor sample expression and the expression in the corresponding estimated control sample all lie in a half-plane. The primary orientation of these displacements, as measured by ν_1, is very similar to the displacement vector between the average tumor cell expression and the average normal cell expression, as shown in the right panel of Figure 3. Thus the large slopes in ν_1 will tend to correspond to genes with highly significant t-tests. However in addition to gene-level information about

differential expression, our procedure also provides measurements of the sample-specific level of differential expression through the coefficients λ_{1j}. A sample with a large value of λ_{1j} will tend to have large expression differences relative to its estimated control sample for the genes that are most differentially-expressed overall. Moreover these differences will tend to be in the same direction as those exhibited by the majority of pairs. A sample with a small value of λ_{1j}, on the other hand, will not tend to have large differences in these genes.

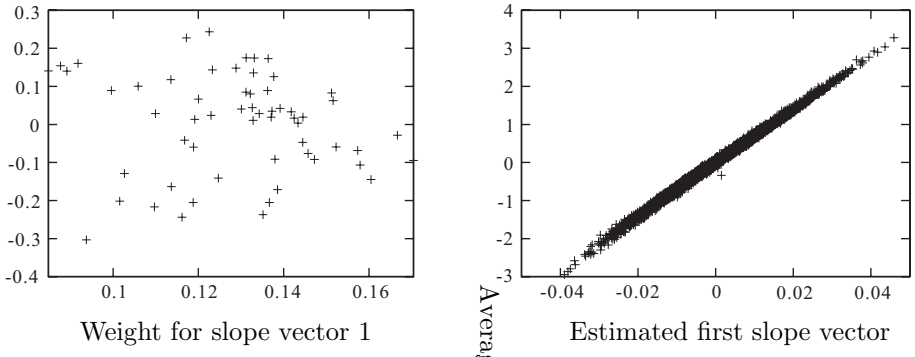

Figure 3. Left panel: Scatterplot of the estimated $(\lambda_{1j}, \lambda_{2j})$ pairs for the breast tumors. Right panel: The slope estimates $\nu_1(i)$ plotted against the displacement vector between the centroid of the tumor samples for gene i and the centroid of the normal samples for gene i.

The primary interest in our method will lie in the slope vectors $\nu_2, \nu_3, \ldots$, and their corresponding weights $\lambda_{2j}, \lambda_{3j}, \ldots$. These are the components that pick up variation in the pairwise tumor/normal displacement vectors that are orthogonal to the average displacement vector, and hence will be mostly missed by t-tests. As an example of how the information in these directions can be extremely important, we consider a clinical aspect of the breast tumor dataset. This study included 40 paired samples obtained from 20 patients, with one sample in each pair taken before chemotherapy treatment, and the other sample taken after the treatment (in no case did the cancer disappear completely, so the tissue samples are classified as tumor both before and after the treatment). In order to determine how the treatment influences gene expression, we located each before/after pair in the left plot of Figure 3, and considered the displacement between the after and before points. For 17 of the 20 pairs, λ_{2j} is greater in the after-treatment sample compared to the before-treatment sample. Thus, the treatment has the effect of displacing samples in the positive direction of ν_2. The influence of the treatment on the λ_{1j} values, however, is not easy to detect (λ_{1j} increases in 8 samples and decreases in 12 samples). In words, there does not seem to be a consistent influence of the chemotherapy on the genes that are most differentially expressed overall. However the genes with large slopes in ν_2 do exhibit a

consistent response to the therapy. We note that it may seem more appropriate to constrain the reconstructed precursor for the two members of a before/after pair to a common value. We did this, and there was no qualitative change in the results (both the 17/20 and the 8/20 findings continued to hold).

At this point we use the breast tumor data to illustrate some aspects of the stability and sampling behavior of our proposed methodology. As stated above, some of the parameters in model (3) are not identified. The subspace spanned by the ν_k is uniquely identified, but the ν_k themselves, and hence the sample-specific coefficients λ_{kj}, are not identified. Therefore it is not easy to discuss the sampling properties of individual parameter estimates. In order to assess whether a single sample may be dominating the slope subspace estimates for the breast cancer data, we fixed $p = 2$ and fit model (3) 59 times, holding out a single sample during each fit. This gives a set of estimates $\{(\nu_1^\ell, \nu_2^\ell), \ell = 1, \ldots, 59\}$ for the 2-dimensional slope subspace. For each ℓ, we computed the canonical angles between $\langle \nu_1^\ell, \nu_2^\ell \rangle$ and $\langle \nu_1, \nu_2 \rangle$, where ν_1 and ν_2 were computed using all 59 tumor samples. We used the largest canonical angle to measure the discrepancy between the two subspace estimates (the largest canonical angle between subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$ will be the supremum of all angles $\cos^{-1}(v_1 \cdot v_2)$ between a unit vector v_1 in $\mathcal{S}_1$ and a unit vector v_2 in $\mathcal{S}_2$). The greatest angle among the 59 leave-one-out sets was 0.06π, where a right angle of 0.5π would be the worst possible value. The median of the 59 values was 0.02π. One implication of this finding is that every linear path in the left side of Figure 3 will be nearly equal to a linear path in the corresponding plot computed when holding out a sample.

A more traditional way to assess sampling variability is to apply a bootstrapping procedure, so that we resample with replacement from the 59 biological samples and re-fit the model. We found that for 95% of the bootstrap samples, the first canonical angle of the $p = 2$ dimensional slope subspace relative to the fit with the the original 59 samples was less than 0.05π, and the second canonical angle was less than 0.31π. From this point of view, one dimension of the slope subspace seems to be quite variable, so it might be valuable to have more samples before drawing any strong conclusions.

3. Conclusion

Our primary goal has been to introduce a new way of approaching differential expression that focuses on the multivariate structure of the displacement vectors between disease samples and paired control samples. A key emphasis is that approaches based on identifying differential expression gene-by-gene, for example through the use of t-tests, correspond to a multivariate perspective focusing solely on the displacement vector between the centroid of the disease samples and the centroid of the normal samples. On the other hand, our approach searches for consistent patterns of discrepancy between normal and disease samples that are

orthogonal to the centroid displacement vector. These directions may be interpreted as determining variant disease development pathways that are consistently followed by a subset of the samples.

The approach relies on two steps. First, paired normal samples are reconstructed by regressing a tumor sample against the observed normal samples. Second, the most important directions for the displacement vectors are identified by carrying out a least-squares fit of a reduced-rank matrix to the matrix of residuals. Both of these steps may fail in certain situations. For example, the available normal samples may not be sufficient to characterize the true range of normal expression behavior. It may also be the case that when a disease sample has experienced a very large shift in expression, the fitted reconstruction will underestimate the magnitude of the true displacement. One potential solution to the second problem would be to identify certain housekeeping genes that are not expected to be associated with the disease process, and use only these genes to estimate the β_N^j. The model also implicitly assumes that most variation in the displacement vectors can be captured in a few dimensions, corresponding to the situation in which continuous expression trajectories $Y(t)$ vary primarily in a low-dimensional subspace. If the displacement vectors appear to have a distribution that is not of reduced dimension, on the other hand, then the method is unable to detect any consistent patterns in the displacement vectors beyond the centroid difference.

Acknowledgement

The author would like to thank Professor Ker-Chau Li and two referees for a number of helpful comments.

References

Lengauer, C., Kinzler, K. W. and Vogelstein, B. (1998). Genetic instabilities in human cancers. *Nature* **396**, 643-649.

Notterman, D. A., Alon, U., Sierk, A. J. and Levine, A. J. (2001). Transcriptional gene expression profiles of colorectal Adenoma, Adenocarcinoma, and normal Tissue examined by Oligonucleotide arrays. *Cancer Research* **61**, 3124-3130.

Perou, C. M., Sorlie, T., Eisen, M., van de Rijn, M., Jeffrey, S., Rees, C., Pollack, J., Ross, D., Johnsen, H., Akslen, L., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S., Lonning, P., Borreson-Dale, A., Brown, P. and Botstein, D. (2000). Molecular portraits of human breast tumors. *Nature* **406**, 747-752.

Vogelstein, B. and Kinzler, K. W. (1993). The multistep nature of cancer. *Trends in Genetics* **9**, 138-141.

Department of Statistics, The University of Michigan, 4062 Frieze Building 105 S. State Street, Ann Arbor, MI 48109-1285, U.S.A.

E-mail: kshedden@umich.edu

(Received March 2001; accepted October 2001)