Biochemistry 218 - BioMedical Informatics 231
Computational Molecular Biology
Final Project

Multivariate Projection Approaches for Microarray Analysis

Gang Yu

Winter 2005

Research on microarrays or gene chips presents a challenge for biologists in functional genomics: because the images generated from microarray experiments are so visually complex that manual comparisons are infeasible, computational tools are required to examine microarrays. For the past several years, there has been an explosion in the numbers of studies on microarray computational tools (Eisen et al., 1998; Burgess, 2001; Risinger et al., 2003; Wouters et al., 2003; Yeung et al., 2004; Saidi et al., 2004; Girolami and Breitling, 2004; Stoyanova et al., 2004; Tan et al., 2004; Busold et al., 2005). In general, four common themes in microarray analysis can be identified. These four themes consist of 1) detection of differential expression, 2) pattern discovery, 3) class prediction, and 4) inference of regulatory pathways and networks (Slonim, 2002).

For the pattern discovery theme, computational tools roughly fall into two major categories. One is multivariate projection methods based upon projections of high-dimensional data in a lower dimensional space and plotting both genes and samples in this lower dimensional space using the biplot (Chapman et al., 2002). This projection into a subspace of low dimensionality can account for the main variance in the data. The other is cluster analysis methods.

This paper discusses approaches in the first category, multivariate projection methods; however, tools in the second category may be mentioned for comparisons. The first part of the paper provides a brief overview of the multivariate methods. Then, algorithms of several major multivariate approaches are presented in the second part. The following part summarizes advantages and drawbacks of multivariate methods with a comparison with cluster tools. The final part is a conclusion.

I. A Brief Overview of Multivariate Projection Approaches

The multivariate projection methods include principal component analysis (PCA), correspondence factor analysis (CFA), spectral map analysis (SMA), partial least squares (PLS) method, and some other variants. Initially, all of these methods were developed in either statistics or other academic areas, but recently used in microarray analysis. Multivariate projection methods help to reduce the complexity (dimensions) of highly dimensional data (n genes versus p samples) and provide means to identify gene patterns or subjects in the data. Projected data are typically displayed in a biplot (genes and samples) in a new space.

PCA is the oldest and best known of the multivariate projection techniques. Historically, PCA dates back to Pearson (1901) and Hotelling (1933). This approach tries to identify components that explain the variance in the data. The central idea of PCA is to reduce the dimensionality or complexity of a data set, while retaining as much as possible of the variation present in the data. The dimension reduction technique is accomplished by introducing a new set of variables "principal components" that are linear combinations of the original variables and uncorrelated to each other. In other words, PCA reproduces the total variance among a large number of variables using a much smaller number of unobservable variables or dimensions called latent factors. Principal components can be determined with different methods such as singular value decomposition (SVD) or some other algorithms. For the just past three years, numerous papers, for example, Peterson, (2003), Barra (2004), Saidi et al., 2004, Tham et al. (2003), Girolami and Breitling (2004), and Hubert and Engelen (2004) have applied this method in microarray studies.

In early 1970s, J. P Benz´ecri developed CFA method for contingency tables and in a sense decomposed the χ^2 statistic. Therefore, distances between objects in CFA have a χ^2 distribution. The method has been widely employed to multivariate data analysis in sociology, environmental science, and marketing research. Kishino and Waddell (2000), Fellenberg et al. (2001), Peterson (2002), Tham et al. (2003), Perelman et al. (2003), Wouters et al. (2003), Tan et al. (2004), and Busold et al. (2005) introduced the method to the investigations of microarray data by displaying the associations between genes and experiments. Since CFA was primarily designed for analyzing contingency tables, it can reveal the association both between and within all the variables (genes and experiments) simultaneously.

Like CFA, SMA was originally developed in 1970s. This method was developed not for biological research either, but for the display of activity spectra of chemical compounds (Lewi, 1976). In the past, SMA has been successfully applied to a wide variety of problems, ranging from pharmacology (Lewi, 1976), virology (Andries et al., 1990), to management and marketing research (Faes and Lewi, 1987). Thielemans et al. (1988) have compared SMA with PCA and CFA, using a relatively small data set from the field of epidemiology. Recently, Wouters et al. (2003) and Peeters et al. (2004) applied this multivariate projection method to microarray analysis. They all argued that SMA would be a promising new tool for microarray data analysis.

PLS is another well-known dimension reduction technique. Wold (1975a, 1975b) developed the PLS approach initially used for modeling information-scarce situations in social science but recently employed in biochemistry. The method relates the data matrix X to a y-response that can be either a single y or multiple Y, i.e., generating a model that predicts y or Y from X. In the computer literature jargon, PLS is known as a supervised method in that it uses both the independent and the dependent variables, whereas PCA is an un-supervised method that considers only independent variables. Datta (2001), Nguyen and Rocke (2002), Park et al. (2002), Johansson et al. (2003), Pérez-Enciso and Tenenhaus (2003), Man et al. (2004), Tan et al. (2004), and Nguyen (2005) have applied this statistical method to microarray data analysis.

II. Algorithms

Due to space limit of this paper, exhaustive explanations of these methods will not be presented; however, a review of the basic elements and major structures of their algorithms is necessary to understand their capabilities in microarray analysis. Concise presentations of the algorithms will be given below.

a) Principal Component Analysis (PCA)

To demonstrate the PCA algorithm, we start to denote $\mathbf{M}_{r\times s}$ as the matrix containing the original expression levels m_{ij} for r genes (rows) in each of s different biological samples (columns). We also define two diagonal matrices with row weights $\mathbf{W}_r$ and column weights $\mathbf{W}_s$, which diagonal elements are the weight coefficients associated with the rows and columns of the matrix **M**. Usually, the weight coefficients are nonnegative. For an unweighted analysis, the weights are obtained by $\mathbf{W}_r$ = diagonal($\mathbf{1}/r$) and $\mathbf{W}_s$ = diag($\mathbf{1}/s$). Alternatively, for weighted analysis the diagonal elements of $\mathbf{W}_r$ and $\mathbf{W}_s$ can be set to appropriate weighting schemes. PCA is characterized by constant weighting of row weights $\mathbf{W}_r$ and column weights $\mathbf{W}_s$.

In PCA, column centering is applied. Centering is defined as a correction of **M** for a mean value to yield the centered matrix **Y**. In column centering, the matrix $\mathbf{Y} = \mathbf{M} - \mathbf{1}_r n_s^T \mathrm{T}_s$ contains deviations from the weighted column means $n_s^T = \mathbf{1}_r^T \mathbf{W}_r^T \mathbf{M}$. In addition, column normalization, or standardization, is employed. In terms of normalization, the original matrix **M** is divided by the square root of the mean sums of squares yielding a normalized matrix **N**. In column normalization, the normalized results is obtained as $\mathbf{N} = \mathbf{MC}_s^{-1}$, with the weighted column-norm $\mathbf{C}_s$, defined as $\mathbf{C}_s$ =diag$((\mathbf{M}^{\mathrm{T}})^2\mathbf{W}_n\mathbf{1}_n)^{1/2}$. The effect of column normalization in the column space is to weight each column dimension proportional to the inverse of its mean sum of squares. Column normalization after column centering is a standard operation in PCA.

The next step is factorization. Factorization of **N** yields factors that are orthogonal to one another, accounting for a maximum of the variance of the data. Matrix **M** is submitted to singular value decomposition (SVD) which transfers it into the product of three matrices **U**, **Λ**, and **V** as shown in equation (2.1):

$$\mathbf{W}_r^{1/2}\mathbf{M}\mathrm{W}_s^{1/2} = \mathbf{U\Lambda V}^{\mathrm{T}} \qquad (2.1),$$

where **U** stands for the eigenvectors of $\mathbf{MM}'$, **V** for the eigenvectors of $\mathbf{M}'\mathbf{M}$, and **Λ** is a diagonal matrix containing the eigenvalues of **M**. **Λ** is an $p\times p$ matrix of singular values, p being the rank of $\mathbf{W}_r^{1/2}\mathbf{MW}_s^{1/2}$. For **U** and **V**, we have $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_p$ and $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}_p$. Consequently, we obtain $(\mathbf{W}_r^{1/2}\mathbf{U})^{\mathrm{T}}\mathbf{W}_r(\mathbf{W}_r^{1/2}\mathbf{U}) = \mathbf{I}_p$ and $(\mathbf{W}_s^{1/2}\mathbf{V})^{\mathrm{T}}\mathbf{W}_s(\mathbf{W}_s^{1/2}\mathbf{V}) = \mathbf{I}_p$.

After the factorization, the final step is projection. With the above matrices, **U**, **Λ**, and **V,** we can project our data into a new space. In the projection, the biplot for PCA is constructed using combinations of two factor-scaling coefficients α and β. We set $\alpha = 1$ and β =1 with symmetric eigenvalue scaling; $\alpha = 1$ and $\beta = 0$ with asymmetric unit column variances; and set $\alpha = 0$, $\beta = 1$ with asymmetric unit row variances respectively. In the new space, the weighted factor scores **S** are obtained from equation (2.1) by $\mathbf{S} = \mathbf{W}_r^{1/2}\mathbf{U\Lambda}^{\alpha}$. We can get factor loadings (i.e., correlations) **L** for FCA by $\mathbf{L} = \mathbf{W}_s^{1/2}\mathbf{V\Lambda}^{\beta}$. An alternative way to get **S** and **F** can be $\mathbf{S} = \mathbf{MW}_s^{1/2}\mathbf{V\Lambda}^{\alpha\text{-}1}$ and $\mathbf{L} = \mathbf{W}_r^{1/2}\mathbf{U\Lambda}^{\beta\text{-}1}$. The latter form is required for positioning supplementary rows or columns by setting their respective weights to zero. These are the major computational features of PCA method for construction of the biplot, which reduces the dimensions of the microarray data.

b) Correspondence Factor Analysis (CFA)

Again, let $\mathbf{M}_{r\times s}$ denote the matrix containing the original expression levels m_{ij} for r genes (rows) in each of s different biological samples or hybridizations (columns). The following mathematical manipulation is to embed both rows and columns of $\mathbf{M}_{r\times s}$ in the same space. The first two or three coordinates of this space contain the bulk of the information on the microarray data. Let ξ_i and ξ_j denote the sum (or the mass) of the ith row and jth column, respectively. By ε , we denote the grand total of **M**. The mass of the jth column, i.e., the sum of row $j,$ is defined as $\psi_j = \xi_j\,/\varepsilon$ *,* and likewise the mass of the ith row is $\theta_i = \xi_i/\varepsilon$. The basis for the calculation is the correspondence matrix **D** with elements $d_{ij} = m_{ij}/\varepsilon$. Then, we derive a matrix **E** with elements $e_{ij} = (m_{ij} - \theta_i\psi_j)/\sqrt{\theta_i\psi_j}$.

Next step relies on the generalized singular value decomposition (SVD) as a factorization method. The generalized SVD of this matrix **E** is defined as $\mathbf{E(S)} = \mathbf{U\Lambda V}^T$. That is, **E** is transferred into the product of three matrices **U**, **Λ**, and **V**. Here again, **U** is the eigenvectors of $\mathbf{EE}'$**,** **V** the eigenvectors of $\mathbf{E}'\mathbf{E}$, and **Λ** is the diagonal matrix containing the eigenvalues of **E**. The elements of **Λ**, ρ_k (k=1,2,…,2n), can be ranked from the largest to the smallest.

Just like in PCA, with the above matrices **U**, **Λ**, and **V,** we project the collected data into a new space. The coordinates for gene i in the new space are then given by $s_{ik} = \rho_k\mu_{ik}/\sqrt{\theta_i}$, for $k = 1,..., J$, where μ_{ik} represents the kth column in **U**. Hybridizations are viewed in the same space with hybridization j given coordinates $t_{ik} = \rho_k\ v_{jk}/\sqrt{\varphi_j}$, for $k = 1,..., J$, where v_{jk} is the kth column in **V**. These coordinates are entitled principal coordinates. In the new space, we plot only the first two or three coordinates. Now the reduction of dimensionality of microarray is achieved.

To make it simple, CFA summarizes high-dimensional data into a low-dimensional space while maintaining the main information by representing the maximum variability in the dataset. The application of CFA to microarray data can help to explore two way intricateness between genes and samples or hybridizations. By combining variables, we can extend CFA to analyzing multiple-table data.

c) Spectral Map Analysis (SMA)

Just like in the computational frameworks for PCA and CFA, we denote $\mathbf{M}_{r\times s}$ as the matrix containing the original expression levels m_{ij} for r genes (rows) in each of s different biological samples (columns). In addition, the row weights $\mathbf{W}_r$ are obtained by $\mathbf{W}_r = \text{diag}(\mathbf{M}1_p/\mathbf{1}_n^T\mathbf{M}1_p)$ and column weights $\mathbf{W}_s$ obtained by $\mathbf{W}_s = \text{diag}(\mathbf{1}_n^T\mathbf{M}/\mathbf{1}_n^T\mathbf{M1}_p)$. The spectral mapping is characterized by constant weighting of row weights $\mathbf{W}_r$ and column weights $\mathbf{W}_s$, or weighting by some properly chosen weighting factor by researchers.

Then, logarithmic reexpression is applied to matrix **M**. That is, data in **M** are transformed to logarithms. **M** with m_{ij} elements is reexpressed as a new matrix **N**, with elements $n_{ij} = \log(m_{ij})$. Logarithmic reexpression allows data in different physical units to be compared to one another. In addition, in many natural systems, changes occur on a multiplicative rather than an additive scale. This reexpression corrects for positive skewness and reduces the effect of large influential values.

The next step is to employ double centering for both row and column to obtain a doubled-centered **P** from the reexpressed matrix **N**. The matrix **P** is obtained by $\mathbf{P} = \mathbf{N} - \mathbf{1}_r n_s^T$ - $n_r^T \mathbf{1}_s^T + \mu \mathbf{1}_n \mathbf{1}_s^T$, where the weighted row means $n_s^T = \mathbf{N}\mathbf{W}_s\mathbf{1}_s$, the weighted column means $n_s^T = \mathbf{1}_r^T \mathbf{W}_r^T \mathbf{M}$, and the global weighted mean $\mu = \mathbf{1}_r^T \mathbf{W}_r^T \mathbf{N}\mathbf{W}_p\mathbf{1}_p$. The double-centering transformation in SMA is symmetric with respect to the rows and columns of the data table. As a result, all absolute aspects of the data are removed. What remains are contrasts between the different rows (genes) and between the different columns (samples) of the data table.

In addition, global normalization is employed. The original matrix **N** is divided by the square root of the mean sums of squares yielding a normalized matrix **Q**. Normalization for the weighted global norm $q = (\mathbf{1}_r\mathbf{W}_r\mathbf{N}^2\mathbf{W}_s\mathbf{1}_s)^{1/2}$ yields the global-normalized matrix. Subsequently, the matrix **Q** is submitted to singular value decomposition (SVD) that transfers it into the product of three matrices, which we denote as **U** (the eigenvectors of $\mathbf{QQ}'$), **V** (the eigenvectors of $\mathbf{Q}'\mathbf{Q}$), and $\mathbf{\Lambda}$ (the diagonal matrix containing the eigenvalues of **Q**):

$$\mathbf{W}_r^{1/2}\mathbf{Q}\mathrm{W}_s^{1/2} = \mathbf{U\Lambda V}^{\mathrm{T}} \qquad (2.2)$$

With the above matrices, **U**, $\mathbf{\Lambda}$, and **V,** we can project our data into a new space. In the new space, the weighted factor scores **S** are obtained from equation (2.2) by $\mathbf{S} = \mathbf{W}_r^{1/2}\mathbf{U\Lambda}^{\alpha}$. We can get factor loadings **L** for FCA by $\mathbf{L} = \mathbf{W}_s^{1/2}\mathbf{V\Lambda}^{\beta}$. For the factor-scaling coefficients α and β in the above equations, researches can select either symmetric scaling with singular values ($\alpha = 0.5, \beta = 0.5$) or asymmetric scaling with unit column variance ($\alpha = 1, \beta = 0$). If the symmetric scaling is used, distances between row points and the correlation structure of the column variables are not fully reproduced. This distortion is most pronounced when the ratios between the eigenvalues ($\mathbf{\Lambda}^2$) associated with the axes of the biplot are very large or very small. In the asymmetric scaling case, only distances between row points are preserved.

d) Partial Least Squares (PLS)

In fact, partial least squares (Wold, 1975) are alternatives to Ordinary Least Square (OLS) in ill-conditioned linear models. Traditional statistical methods such as OLS for classification do not work when there are more variables than there are samples. Gene expression data from microarrays characteristically have many measured variables (genes) and only a few observations (experiments). Thus PLS can be a feasible option.

PLS method estimates an orthogonal basis for the covariates, which depends on both the covariate and response value. Several equivalent algorithms for computing partial least-squares have been published in the literature. Here, a "classical" or traditional version of the algorithms is presented. Actually, the essence of all of these PLS algorithms are the same. That is, the development of partial least squares was motivated by a representation of the standardized response vector **Y** and the predictor matrix **X.**

The algorithm is briefed as follows. PLS is developed based on the regression between the scores for the two variable matrices, i.e., the descriptor or independent variable matrix **X** and the response or dependent variable matrix **Y.** In microarray analysis, for instance, the independent variable can be genes or expression levels, while the dependent variables may be some specific phenotypes or cases. This method is based on the projection of the original multivariate data matrices **X** and **Y** down onto smaller matrices that hold the coordinates of the new axes. Mathematically, this projection is realized by resolving **X** into the product of smaller matrices, **T** (the scores matrix, which holds the coordinates of the new axes) and **P** (the **X**-loading matrix, which contains the directions of the axes and shows the influence of the *x* in each component), and similarly, **Y** into the product of **U** (the scores matrix) and **Q** (the **Y**-loading matrix), as shown below in equations (2.3), (2.4), and (2.5). This model can be considered as consisting of outer relations (**X** and **Y** individually) and an inner relation (2.5) which links the two variable matrices.

$$\mathbf{X} = \mathbf{TP} + \mathbf{E} \text{ (the outer relation)} \qquad (2.3),$$

$$\mathbf{Y} = \mathbf{UQ} + \mathbf{F} \text{ (the outer relation)} \qquad (2.4),$$

and

$$\mathbf{U} = \mathbf{T} + \mathbf{H} \text{ (the inner relation)} \qquad (2.5),$$

where matrices **E**, **F**, and **H**, entitled "residual" matrices, contain the model error and random noise.

Moreover, the PLS calculations can introduce an auxiliary matrix **W** (PLS weights) that expresses the correlation between **U** and **X** and is used to calculate **T**. The PLS model can also be constructed using the singular value decomposition (SVD). With the assistance of PLS, **X** and **Y** are replaced by **T** and **U** which have better properties, i.e., orthogonality, and which also span the multidimensional vector space of **X** and **Y**, respectively. Hence, the intention is to describe **X** and **Y**, as well as possible by making $\|E\|$ and $\|F\|$ respectively, as small as possible, and at the same time obtain a useful relationship between **X** and **Y**. $\|E\|$ and $\|F\|$ are length or norm of **E** and **F**, defined by $\|E\| = \sqrt{\mathrm{E} \bullet \mathrm{E}} = \sqrt{e_1^2 + e_2^2 + ... + e_n^2}$ and $\|F\| = \sqrt{\mathrm{F} \bullet \mathrm{F}} = \sqrt{f_1^2 + f_2^2 + ... + f_n^2}$ in linear algebra, where $e_1, e_2, ..., e_n$, and $f_1, f_2, ..., f_n$ are entries of E and F, respectively. Via the projection, data reduction is achieved.

III. Strength and Limitation of Multivariate Projection Methods

Both of the predominant merit and drawback of the multivariate projection methods are simultaneously embedded in their algorithms: reduction of data dimensions. This reduction simplifies the overwhelming data obtained from microarray experiments, while the simplification may omit some important biological links or properties of gene expression. Another merit of multivariate analysis is that these models can handle data when information (for example, some gene expression) is missing. Projection methods generally aim at explaining the major trends in the data while ignoring minor fluctuations. Nonetheless, the treatment on missing information and minor fluctuations can lower the sensibility of the models.

Another limitation of some of these methods is that their algorithms apply logarithmic re-expression. Consequently, contrasts at a less reliable level of gene expression are considered to be of equal importance to contrasts at a more reliable level (Peeters et al., 2004). A further limitation is that some data assumptions underlying these multivariate methods highlights their lack of biological validity (Girolami and Breitling, 2004). Pérez-Enciso and Tenenhaus (2003) argued that there is no guarantee that dimension reduction techniques provide researchers with a fully meaningful biological response. Consequently, it is difficult to make biological interpretation of some of the results obtained via these methods. Moreover, except for a couple of PLS applications, existing investigations using these approaches have been mainly limited to small data sets. There can be other challenges if these methods are employed to large data sets.

Contrast to multivariate projection approaches, cluster analysis methods, which include agglomerative hierarchical clustering analysis (HCA), self-organizing map (SOM), and support vector machines (SVM), are more widely used. However, widely applications do not mean that cluster methods are superior to multivariate projection approaches. The cluster analysis methods have their own limitations. First, these methods produce results that are highly dependent on the distance measure and clustering techniques that are used. Multivariate methods are less dependable on measures and the techniques used. A second limitation is that clustering methods, especially hierarchical clustering, are sensitive to datum noise (Segal, 2003), whereas multivariate projections models are much less sensitive to data noise. In reality, one of the key statistical concepts highlighted by the microarray experiment is that data are inherently noisy and collinear, and that randomness is inherent in any sampling process (Bergeron, 2003). From this perspective, multivariate methods have an advantage. Third, the clustering is rather local than global (Segal, 2003). Fourth, in cluster models, each gene has to be assigned to only one cluster, but some genes could have fallen into two or more clusters. Multivariate approaches are relatively flexible. Finally, conventional clustering methods only allow for classification of either genes or biological samples alone, but do not allow interpretations for the association between genes and samples (Wouters et al., 2003). In some sense, multivariate methods seem to have the same limitation.

Essentially, both multivariate projection and cluster methods are "simplification" approaches for tremendous and complex microarray data but via different channels. The projection algorithms achieve the goal by "reduction," while cluster ones reach the objective by "grouping" or "classification."

IV. Conclusion

For microarray analysis, the good news is that there is a wide range of approaches available. However, too many tools can confuse researchers. The appropriate choice of tools depends both on the data available and on the goals of the research. In microarray studies, if resources allow, researchers may consider using some approaches together. This point has already been seen in the microarray literature. Peterson (2002) argued that a certain multivariate method could augment cluster analysis in the search for unique expression profiles among named genes or ESTs. Some others indicated in their investigations that multivariate projection approaches and cluster tools might supplement to each other.

At present, microarray analysis tools still have serious limitations and no single method, or even a set of algorithms, can be recommended to exclusion of others (Meltzer, 2001; Wouters et al., 2003). Unless a brand-new algorithm being developed, there is no one-size-fits-all solution currently and very possibly in the near future, due to the limitations inherent in their algorithms by nature. Although there has been noticeable progress in developing statistical methods to handle microarray data, the search for new methods and improvement of existing tools will still be a subject of active investigation in the future.

References:

Barra V. (2004). Analysis of gene expression data using functional principal components. *Computer Methods and Programs in Biomedicine*. 75, 1-9.

Baxevanis D. and Ouellette B. (2001). Bioinformatics: a practical guide to the analysis of genes and proteins. New York: Wiley-Interscience. 297.

Bayani J., Brenton J.D., Macgregor P.F., Beheshti B., Albert M., Nallainathan D., Karaskova J., Rosen B., Murphy .J, Laframboise S., Zanke B., and Squire J.A. (2002). Parallel analysis of sporadic primary ovarian carcinomas by spectral karyotyping, comparative genomic hybridization, and expression microarrays. *Cancer Res*. 62(12), 3466-3476.

Bergeron B. (2003). Bioinformatics computing. Upper Saddle River, NJ: Prentice Hall. 260-283.

Bittner M., Meltzer P., and Trent J. (1999). Data analysis and integration: of steps and arrows. *Nat. Genet.* 22, 213-215.

Burgess J.K. (2001). Gene expression studies using microarrays. *Clin. Exp. Pharmacol. Physiol.* 28, 321-328.

Busold C.H., Winter S., Hauser N., Bauer A., Dippon J., Hoheisel J.D., and Fellenberg K. (2005). Integration of GO annotations in Correspondence Analysis; facilitating the interpretation of microarray data. *Bioinformatics*. [Epub ahead of print].

Chapman S., Schenk P., Kazan K., and Manners J. (2002). Using biplots to interpret gene expression patterns in plants. *Bioinformatics.* 18, 202–204.

Datta S. (2001). Exploring relationships in gene expressions: a partial least squares approach. *Gene Expr*. 9(6), 249-55.

Eisen M.B., Spellman P.T., Brown P.O. and Botstein D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA.*, 95(25), 14863-14868.

Fellenberg K., Hauser N., Brors B., Neutzner A., Hoheisel J., and Vingron M. (2001). Correspondence analysis applied to microarray data. *Proceedings of the National Academy of Sciences U.S.A*. 98, 10781–10786.

Gabriel K.R. (1971). The biplot graphical display of matrices with applications to principal component analysis. *Biometrika.* 58, 453–467.

Gibas C. and Jambeck P. (2001). Developing Bioinformatics Computer Skills. Sebastopol, CA: O'Reilly. 37 and 311.

Girolami M., and Breitling R. (2004). Biologically valid linear factor models of gene expression. *Bioinformatics*. 20(17), 3021-33.

Hotelling H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology.* 24, 417–441.

Huang J. and Harrington D. (2004). Iterative partial least squares with right-censored data analysis: a comparison to other dimension reduction techniques. *Biometrics*. 61(1), 17-24.

Landgrebe J., Welzl G., Metz T., van Gaalen M. M., Ropers H., Wurst W., and Holsboer F. (2002). Molecular characterisation of antidepressant effects in the mouse brain using gene expression profiling. *Journal of Psychiatric Research*. 36, 119–129.

Hubert M, Engelen S. (2004) Robust PCA and classification in biosciences. *Bioinformatics*. 20(11), 1728-36.

Johansson D., Lindgren P., Berglund A. (2003). A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription. *Bioinformatics*. 19(4), 467-473.

Kishino H., Waddell P.J. (2000). Correspondence analysis of genes and tissue types and finding genetic links from microarray data. *Genome Inform Ser Workshop Genome Inform*. 11, 83-95.

Lewi P.J. and Moereels R. (1994). Receptor mapping and phylogenetic clustering. In Advanced computer-assisted techniques of drug discovery, H. van de Waterbeemd (ed), 131–162. Weinheim, Germany: VCH.

Man M.Z., Dyson G., Johnson K., Liao B. (2004). Evaluating methods for classifying expression data. *J Biopharm Stat*. 14(4), 1065-1084.

Meltzer P.S. (2001). Large scale genome analysis. In Bazevanis A.D. and Ouellette B.F.F. (eds.) Bioinformatics: A practical guide to the analysis of genes and proteins. New York: Wiley-Interscience.

Mount D. (2002). Bioinformatics: sequence and genome analysis. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press. 519.

Nguyen D.V. (2005). Partial least squares dimension reduction for microarray gene expression data with a censored response. *Math Biosci*. 193(1), 119-137.

Nguyen D.V., Rocke D.M. (2002a). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*. 18(1), 39-50.

Nguyen D.V., Rocke D.M. (2002b). Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinformatics*. 18(12), 1625-1632.

Park P.J., Tian L., Kohane I.S. (2002). Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*. 18 (Suppl 1), S120-S127.

Pearson K. (1901). On lines and planes of closest fit to points in space. *Philosophical Magazine*. 2, 559–572.

Peeters P.J., Gohlmann H.W., Van den Wyngaert I., Swagemakers S.M., Bijnens L., Kass S.U., Steckler T. (2004). Transcriptional response to corticotropin-releasing factor in AtT-20 cells. *Mol Pharmacol*. 66(5),1083-1092.

Perelman S., Mazzella M., Muschietti J., Zhu T., and Casal J.J. (2003). Finding unexpected patterns in microarray data. *Plant Physiol*. 133(4), 1717-1725.

Pérez-Enciso M. and Tenenhaus M. (2003). Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. *Hum Genet* 112, 581–592.

Peterson L.E. (2003). Partitioning large-sample microarray-based gene expression profiles using principal components analysis. *Computer Methods and Programs in Biomedicine* 70, 107–119.

Peterson L.E. (2002). Factor analysis of cluster-specific gene expression levels from cDNA microarrays. *Computer Methods and Programs in Biomedicine.* 69(3). 179-188.

Risinger J.I., Maxwell G.L., Chandramouli G.V., Jazaeri A., Aprelikova O., Patterson T., Berchuck A., and Barrett J.C. (2003). *Cancer Res.*, 63, 6̃11.

Samir A Saidi, Cathrine M Holland, David P Kreil, David J C MacKay, D Stephen Charnock-Jones, Cristin G Print and Stephen K Smith. (2004). Independent component analysis of microarray data in the study of endometrial cancer. *Oncogene*. 23(39), 6677-6683.

Slonim D.K. (2002). From patterns to pathways: gene expression data analysis comes of age. *Nat Genet.* 32. S502–S508.

Querec T.D., Stoyanova R., Ross E., Patriotis C. (2004). Normalization of single-channel DNA array data by principal component analysis. *Bioinformatics*. 20(11), 1772-1784.

Segal E. (2003). Class notes for BMI 231/BIOC 218: Computational Molecular Biology. Stanford University.

Tavazoie S., Hughes J.D., Campbell M.J., Cho R.J., and Church G.M. (1999). Systematic determination of genetic network architecture. *Nature Genetics.* 22, 281–285.

Tan Q., Brusgaard K., Kruse T.A., Oakeley E., Hemmings B., Beck-Nielsen H., Hansen L., Gaster M. (2004). Correspondence analysis of microarray time-course data in case-control design. *J Biomed Inform*. 37(5), 358-365.

Tan Y., Shi L., Tong W., Hwang G.T., Wang C. (2004). Multi-class tumor classification by discriminant partial least squares using microarray gene expression data and assessment of classification models. *Comput Biol Chem*. 28(3), 235-244.

Tham C.K., Heng C.K., Chin W.C. (2003) Predicting risk of coronary artery disease from DNA microarray-based genotyping using neural networks and other statistical analysis tool. *J Bioinform Comput Biol*. 1(3), 521-539.

T¨or¨onen P., Kolehmainen M., Wong G., and Castr´en E. (1999). Analysis of gene expression data using self organizing maps. *FEBS Letters.* 451, 142–146.

Wold H. (1975a). Perspectives in probability and statistics. London: Academic Press.

Wold, H. (1975b). Soft modeling by latent variables: The nonlinear iterative partial least squares (NIPALS) approach. In *Perspectives in Probability and Statistics, in Honor of M.S.*, Bartlett, J. Gani (ed), 117–142. New York: Academic Press.

Wold S., Ruhe A., Wold H. and Dunn III, W.J. (1984) The collinearity problem in linear regression. the partial least squares approach to generalized inverses. *SIAM J. Sci. Stat. Comput.* 5, 735–743.

Yeung L.K., Szeto L.K., Liew A.W., Yan H. (2004) Dominant spectral component analysis for transcriptional regulations using microarray time-series data. *Bioinformatics*. 20(5), 742-749.