12-23-2004

Clustering and Classification Methods for Gene Expression Data Analysis

Elizabeth Garrett-Mayer

The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University & Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, esg@jhu.edu

Giovanni Parmigiani

The Sydney Kimmel Comprehensive Cancer Center, Johns Hopkins University & Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, gp@jimmy.harvard.edu

Clustering and Classification Methods for Gene Expression Data Analysis.

Elizabeth Garrett-Mayer[1,2] *Giovanni Parmigiani*[1,2,3]

1. The Sidney Kimmel Comprehensive Cancer at Johns Hopkins University

2. Department of Biostatistics, Johns Hopkins University

3. Department of Pathology, Johns Hopkins University

December 23, 2004

1 Introduction

Efficient use of the large data sets generated by gene expression microarray experiments requires computerized data analysis approaches (1; 2). In this chapter we briefly describe and illustrate two broad families of commonly used data analysis methods: class discovery and class prediction methods. Class discovery, also referred to as clustering or supervised learning, has the goal of partitioning a set of objects (either the genes or the samples) into groups that are relatively similar, in the sense that objects in the same group are more alike than objects in different groups (3; 4). A typical application is to generate hypotheses about novel disease subtypes (5; 6). Class prediction, also referred to as classification or supervised learning, has the goal of determining whether an object (usually a sample, but sometimes a gene) belongs to a certain class (7; 8). A typical application is classification of patients into existing disease subtypes or prognostic classes (9; 10) using gene expression information.

In our discussion, “sample” refers generically to any type of biological material that is processed and hybridized to a chip. For example, in a study of breast cancers, the samples could represent the breast cancer tissues biopsied from a group of women. “Gene” is used loosely to refer to the features on the arrays, such as sequences from genes or ESTs, single oligonucleotides in Agilent arrays, oligonucleotide sets in Affymetrix arrays and so forth. “Object” refers to the entity being clustered, and can be either a gene or a sample, as the same algorithms can often be applied symmetrically to both. “Attribute” is any feature of the object being clustered. If we cluster samples, genes are typically at-

tributes, and vice versa. “Phenotype” refers to any clinical or biological characteristic of a sample or the person or organism from which the sample is associated, such as disease subtype, age, gender, or time to disease progression.

To demonstrate the clustering methods in this chapter, we use a gene expression microarray dataset published by Hedenfalk and colleagues (11) and including samples from twenty-two breast cancers, of which seven are from patients with known BRCA1 mutations, eight from patients with known BRCA2 mutations, and seven are sporadic. Complementary DNA (cDNA) was obtained from each tumor sample and hybridized to two channel cDNA arrays which included spots for 3226 genes and ESTs. The reference sample was cell-line MCF-10, a nontumorigenic breast cell line. Data from this study is available at `http://www.nhgri.nih.gov/DIR/Microarray`.

Statistical computing environments typically offer a rich set of alternatives for clustering and classification. In particular the free and open source computing environment R (12) and the associated Bioconductor (13) project cover most standard tools, a wide variety of developmental tools and offer the flexibility for implementing custom solutions. A range of free and open source tools can be accessed via the website `www.arraybook.org`. The site `http://ihome.cuhk.edu.hk/~b400559/arraysoft.html` maintains a catalog of both free and commercial microarray data analysis software.

2 Protocols

2.1 Clustering

Clustering techniques can be used in microarray analysis to a) facilitate visual display and interpretation of experimental results; and b) suggest the presence of subgroups of objects (genes or samples) that behave similarly. The input of a cluster analysis are the gene expression values of the samples in an experiment, with no additional phenotype information. Depending on the approach, the output can be a list of subgroups, or a visualization that simplifies manually establishing subgroups. In some applications, unsupervised methods are used even though phenotype information is available. The goal is often to see how the clusters of samples that arise from an unsupervised approach compare to the known phenotypes.

2.1.1 Distance and Similarity

To determine which objects cluster together, we must have a way of measuring how similar, or dissimilar any two of them are. Most clustering approaches will allow as input a matrix whose entries measure similarity, or dissimilarity, between each pair objects. Choosing this measure is one of the most critical, yet often underappreciated, aspects of a cluster analysis. Different measures reflect different goals, and thus can have a strong influence on the resulting clusters. Here we discuss in detail three: the correlation coefficient, which will bring together objects whose patterns of change are similar; the Eu-

clidean distance, which will bring together objects whose absolute expressions are similar, and the uncentered correlation, which achieves a compromise between the previous two.

The Pearson correlation coefficient measures the strength of a linear association between the expression levels of objects. In the case of genes j and k, it is defined by

$$\rho_{jk} = \frac{\sum_{s=1}^{S}(x_{sj} - \bar{x}_j)(x_{sk} - \bar{x}_k)}{\sqrt{\sum_{s=1}^{S}(x_{sj} - \bar{x}_j)^2 \sum_{s=1}^{S}(x_{sk} - \bar{x}_k)^2}} \tag{1}$$

where x_{sj} is the gene expression for gene j in sample s and $\bar{x}_j$ is the average gene expression of gene j across all samples. A symmetric definition applies to the correlation between samples. The correlation takes values ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation). A correlation of 0 means that there is no linear relationship between the two genes. For analyses that require positive similarity matrices, it is common to use the absolute value of the correlation with the rationale that high negative and high positive correlations both may imply an underlying common mechanism. The correlation coefficient is unitless, but is sensitive to nonlinear transformation of the data, such as the logarithm. For non-linear relationships, the correlation coefficient may not adequately describe similarity. Another drawback is that it may be sensitive to noise.

The Euclidean distance measures geometric distance between two objects. In the case of genes j and k, it is defined by

$$d_{jk} = \sqrt{\sum_{s=1}^{S}(x_{sj} - x_{sk})^2}. \tag{2}$$

A symmetric definition applies to the correlation between samples. It takes values from 0 to ∞ and it retains the units of the input gene expression measurements. It grows with the number of samples included in the dataset.

The uncentered correlation (14) is similar to the Pearson correlation but is evaluated without centering :

$$\epsilon_{jk} = \frac{\sum_{s=1}^{S} x_{sj} x_{sk}}{\sqrt{\sum_{s=1}^{S} x_{sj}^2 \sum_{s=1}^{S} x_{sk}^2}}. \quad (3)$$

As the Pearson correlation, this is unitless, but is sensitive to absolute magnitudes as the Euclidean distance. As a result it will be less likely to be influenced by genes whose variation is mostly noise.

For a summary of other distance and similarity metrics, see (15).

2.1.2 Hierarchical Clustering

Hierarchical clustering is used to partition objects into a series of nested clusters (5; 6), by contrast with approaches that find a single partition (16). To illustrate, a hierarchical clustering analysis of both genes and samples in the Hedenfalk data is shown in Figure 1, along with a grey scale image of gene expression levels. The similarity used is the uncentered correlation. The hierarchy of clusters of samples is displayed using a tree-like structure called dendrogram. Dendrograms join objects, or clusters of objects, to form increasingly large clusters. The height at which two clusters are joined represents how similar they are, with low heights representing high similarity. Samples in Figure 1 are labeled by their type (BRCA1, BRCA2, or sporadic), though these types are not used in

constructing the dendrogram.

There are two kinds of hierarchical clustering approaches: agglomerative and divisive. The agglomerative approach begins by assuming that each object belongs to its own separate cluster. At the first step, the two most similar objects are combined to form a new cluster. Then the next most similar clusters or object are combined and so forth. This is a bottom-up approach in the sense that the clustering starts at the bottom of the dendrogram of Figure 1 and works its way up until all objects belong to one cluster. As part of the agglomerative approach, it is necessary to specify a linkage method, that is a way of defining similarity of clusters based on similarities of cluster members. Some of the commonly used linkage methods are single, average, and complete in which clusters are linked based on the similarity of the closest members, the average similarity, and the similarity of the furthest members.

The divisive approach works from the top of the dendrogram, where all objects belong to one cluster. At the first step, it finds the best division of the objects so that there is the highest similarity among objects within clusters and the most dissimilarity between clusters. This process continues, where the best cluster partition is chosen at each step until all objects are in their own clusters. Details of hierarchical clustering can be found in (4).

An important consideration when applying or interpreting hierarchical clustering results is that there is not a unique dendrogram for a given hierarchical clustering result. For each split in a dendrogram, it is arbitrary which branch is drawn to the right or left, and users

need to specify criteria for this choice. As such, many dendrograms can be drawn for a given hierarchical clustering result and closeness of objects should be based on the height at which they are joined, rather than their ordering in the dendrogram.

Preselection of genes can significantly affect clustering of samples and vice versa. Selecting genes that show at least a certain amount of variation across samples is useful to reduce the sensitivity of clustering results to noise variation. Selecting genes whose variation is associated with a phenotype of interest is also common, though when that is done the correspondence of clusters to phenotype cannot be invoked as validation of the clustering results, as the correspondence will be inflated by the preselection. To illustrate, compare the left panel of Figure 1, which includes all genes in the experiment, to the right panel, where only the top 25% of genes associated with the BRCA types are included. The dendrogram on the left has short branch links and cascading patterns, both of which weaken the case for the existence of clusters. None of the main partitions has any relation to the BRCA type. On the right, the branch links at the top are longer and there is some evidence of two major clusters, which separate well the BRCA1 from the BRCA2 cases. While in general a correspondence between clusters found by unsupervised analyses and sample phenotypes can be taken as independent supporting evidence of the existence of clusters of biological significance, in this case this argument would be circular, because the sample phenotypes were used in selecting the genes for clustering.

2.1.3 K-means Clustering and Self-Organizing Maps

K-means clustering (17) partitions objects into groups that have little variability within clusters and large variability across clusters. The user is required to specify the number k of clusters a priori. Estimation is iterative, starting with a random allocation of objects to clusters, re-allocating to minimize distance to the estimated "centroids" of the clusters, and stopping when no improvements can be made. The centroid is the point whose attributes take the mean expression level of the objects in the clusters. K-medoids clustering is similar, except that the center of the clusters is defined by "medoids", similar to centroids, but based on medians (4). Specification of k can be difficult, though there are ways of gaining insight into the appropriate number of clusters, such as using principal components analysis. A closely related approach is that of self-organizing maps (18; 7; 15), now common in in gene expression data (16).

2.1.4 Principal Components Analysis and Multi-Dimensional Scaling

Principal Components Analysis (PCA) (19; 20; 21) and Multidimensional Scaling (MDS) are techniques whose goal is to reduce the dimensionality of data to facilitate visualization and additional analysis. They are often used as a preliminary step to the clustering of large data sets and are commonly applied to gene expression data (22; 23; 24; 25; 26; 27; 28).

PCA creates summary attributes, or "components", that are weighted averages of the original attributes, are uncorrelated to each other, and are such that most of the variability

in the data is concentrated in few components. During the estimation process, as many components as there are attributes are calculated. Users select a small number, chosen to retain a sufficient fraction of the variability. These are often plotted to visually search for clusters. A strength of PCA is that redundant information is represented in a single component, while a drawback is that the components may lack clear biological interpretations.

The first three PC's for the Hedenfalk data are shown in Figure 2. Here, instead of having to visualize thousands of genes per sample, we here use three weighted averages of genes. Together, they describe 38% of the variability in the data. The samples appear to cluster in subgroups. When phenotype information is available, one can check putative subgroups again the phenotype information, or gauge how the variability in expression relates to the variability in phenotypes. For example, in the top-left panel, the sporadic samples tend to have high values for component 2 and relatively low values for component 1. BRCA2 samples are distributed differently with most having either very low values for component 2 or high values for both components 1 and 2. The four areas in the plot created by the two intersecting lines are discriminating between different BRCA types. The results for components 1 versus 3 and 2 versus 3 also show some clustering, though these are not as clearly related to BRCA types.

MDS starts from a distance matrix between objects and finds the locations of these objects in a low dimensional space that best preserves the original distances. For example, given objects in three dimensions, MDS may find the two dimensional map of these ob-

jects that is most faithful to the original three-dimensional distances. The result is similar to the PCA result: we have summary variables, the coordinates of the map, that describe a large fraction of the variability in the gene expression measures, and that can be visually inspected to identify clusters. Two examples of MDS as applied to gene expression data can be found in (29; 30).

2.1.5 Limitations of cluster analysis

Clustering techniques for high dimensional data are exploratory. Their strength is in providing rough maps and suggesting directions for further study. Substantial additional work is necessary to provide context and meaning to groups found by automated algorithms. This includes cross referecing of existing knowledge about genes and samples as well as additional biological validation.

Clustering results are sensitive to a variety of user-specified inputs. The clustering of a large and complex set of objects can, like arranging books in a collection, be planned in different ways depending on the goals. From this perspective, good clustering tools are responsive to users' choices, not insensitive to them, and sensitivity to input is a necessity of cluster analysis rather than a weakness. This also means, however, that use of a clustering algorithm without knowledge of its workings, the meaning of inputs, and their relationship to the biological questions of interest is likely to yield misleading results.

Clustering results are generally sensitive to small variations in the samples and the genes chosen and to outlying observations. This means that a number of the data-analytic

decisions made during normalization, filtering, data transformations, and so forth will have an effect on results. When conclusions drawn from clustering go beyond simple data visualization, it is important to provide accurate assessments of the uncertainty associated with the clusters found. Uncertainty from sampling and outliers can be addressed within model-based approaches (31) or alternatively using resampling techniques (32; 33; 34). The consequences of choosing among plausible alternative transformations, normalizations, and filtering should be addressed by sensitivity analysis, that is by repeating the analysis and reporting conclusions that are consistent across analyses.

2.2 Classification

Classification techniques can be used in microarray analysis to predict sample phenotypes based on gene expression patterns. While novel and microarray specific classification tools are constantly being developed, the existing body of pattern recognition and prediction algorithms provide effective tools (35). Dudoit and colleagues (36) offer a practical comparison of methods for the classification of tumors using gene expression data. Relevant tools from the statistical modeling tradition include: discriminant analysis (37), including linear, logistic, and more flexible discrimination techniques; tree-based algorithms, such as classification and regression trees (CART) by (38) and variants; generalized additive models (39); and neural networks (40; 7; 41). Appropriate versions of these methods can be used for both classification and prediction of quantitative responses such as continuous measures of aggressiveness. Some of these methods are briefly reviewed

here.

2.2.1 Dimension Reduction

Because of the large number of genes that can be used as potential predictors, it is useful to preselect a subset of genes, or composite variables, likely to be predictive and then investigate in depth the relationship between these and the phenotype of interest. For example, genes with nearly constant expression across all samples can be eliminated. Additional screening can be based on measures of marginal association, such as the ratio of within-group variation to between-group variation, or the measure used in (42), though these can miss important genes that act in concert with others but have no strong marginal effects.

Parsimonious representations of the data may be identified when there is knowledge of important pathways that can be used to manually construct new and more highly explanatory variables. When such knowledge is not available we need to apply discovery techniques such as those described earlier; for example, the centroids of clusters or the variables identified by PCA can be used as predictors. Composite variables that are easily measurable and interpretable in terms of the original gene expression are generally preferable. Automatic approaches for preclustering variables before classification are also useful (43).

2.2.2 Evaluation of Classifiers

Classifiers based on gene expression are generally probabilistic, that is they only predict that a certain percentage of the individuals that have a given expression profile will also have the phenotype, or outcome, of interest. Therefore, statistical validation is necessary before models can be employed, especially in clinical settings (44; 45).

The most satisfactory approaches to validation require the use of data other than those used to develop the classifier. When only a single study is available, this can often be achieved by setting aside samples for validation purposes, as illustrated by (36). Statistical validation of probabilistic models (46) should focus on both refinement, that is the ability of the classifier to discriminate between classes, and calibration, that is, the correspondence between the fraction predicted and the fraction observed in the validation sample.

An alternative to setting aside samples for validation is the so-called cross-validation. For example, K-fold cross-validation consists of splitting the data in K subsets, and training the classifier K times, setting aside each subset in turn for validation. The average classification rates in the K analyses is then an unbiased estimate of the correct classification rate (47).

A potentially serious mistake is to evaluate classifiers on the same data that were used for training. When the number of predictors is very large, a relatively large number of predictors will appear to be highly correlated with the phenotype of interest as a result of the random variation present in the data. These spurious predictors have no biological

foundation and do not generally reproduce outside of the sample studied. As a result, evaluation of classifiers on training data tends to give overly optimistic assessments of validity. In plausible settings, classifiers can appear to have a near perfect classification ability in the training set without having any biological relation with phenotype (48). All aspects of learning a classifier need to be properly cross-validated to avoid inflated estimates of performance.

2.2.3 Predictive Analysis of Microarrays (PAM)

A straightforward approach to classification is the nearest centroid classifier. This computes, for each class, a centroid given by the average expression levels of the samples in the class, and then assigns new samples to the class whose centroid is nearest. This approach is similar to k-means clustering except clusters are now replaced by known classes. With a large number of genes this algorithm can be sensitive to noise. A recent enhancement uses shrinkage: for each gene, differences between class centroids are set to zero if they are deemed likely to be due to chance. This approach is implemented in the Prediction Analysis of Microarray, or PAM (49) software. Shrinkage is controlled by a threshold below which differences are considered noise. Genes that show no difference above the noise level are removed. A threshold can be chosen by cross-validation, as shown in Figure 3 for the Hendefalk data. High thresholds, on the right, include few genes, and lead to classifiers that are prone to errors. As the threshold is decreased more genes are included and estimated classification errors decrease, until they reach a

bottom and start climbing again as a result of noise genes —a phenomenon known as overfitting.

2.2.4 Top scoring pairs

Another simple and very effective tool is the top-scoring pair(s), or TSP, classifier (50). In a two-class classification, this looks for pairs of genes such that gene 1 is greater than gene 2 in class A and smaller in class B. This handles effectively issues of normalization as the pair provides an internal control and is likely to give generalizable results. TSP classifiers are transparent and interpretable and provide specific hypotheses for follow-up studies. In cancer data the TSP classifier achieves prediction rates that are as high as those of alternative approaches which use considerably more genes and complex procedures (50).

2.2.5 Nearest-Neighbor Classifiers

Nearest-neighbors classifiers (51), assign samples to classes by matching the gene expression profile to that of samples whose class is known. A simple implementation is to choose a rule for finding the k nearest neighbors and then deciding the classification by majority vote. Nearest-neighbor classifiers are robust, simple to interpret and implement, and do not require, although they may benefit from, preliminary dimension reduction. Nearest-neighbor algorithms are also used in several packages for imputation of missing data.

2.2.6 Support Vector Machines

Support vector machines (SVMs) (52) seek cuts of the data that separate classes effectively, that is by large gaps. Technically, SVMs operate by finding a hypersurface in the space of gene expression profiles, that will split the groups so that there is the largest distance between the hypersurface and the nearest of the points in the groups. More flexible implementations allow for imperfect separation of groups. See (53) and (54) for details of SVMs and generalizations, while (55) and (56) give examples of analysis of gene expression data using SVMs.

2.2.7 Discriminant Analysis

Discriminant analysis (57) and its derivatives are approaches for optimally partitioning a space of expression profiles into subsets that are highly predictive of the phenotype of interest, for example by maximizing the ratio of between-classes variance to within-class variance. (7) and (21) give details, while (58) discusses flexible extensions of discriminant analysis (FDA) and (59) provides a discussion of discriminant analysis in the context of gene expression array data.

2.2.8 Classification Trees

Classification trees recursively partition the space of expression profiles into subsets that are highly predictive of the phenotype of interest (38). They are robust, easy-to-use, and can automatically sift large data sets, identifying important patterns and relationships. No

prescreening of the genes is required. The resulting predictive models can be displayed using intuitive graphical representations. An example in which classification trees have been applied to gene expression data can be found in (60).

2.2.9 Regression-based Approaches

Linear models, generalized linear models, generalized additive models and the associated variable selection strategies provide standard tools for selecting useful subset of genes and developing probabilistic classifiers. A limitation of these techniques is that they cannot generally handle more genes than there are samples. This can be circumvented using forward selection approaches that progressively add genes to the classifier. Recent, more accurate approaches are based on the so-called stochastic search methods (61), that generate a sample of plausible subsets of explanatory variables. The selected subsets are then subjected to additional scrutiny to determine the most appropriate classification algorithm. A combination of stochastic search with principal component analysis and other orthogonalization techniques has proven effective in high-dimensional problems (62; 63), and has recently been employed in microarray data analysis (23).

2.2.10 Probabilistic Model-based Classification

Model based classification is based on the specification of a probability distribution that describes the variability of the expression values. Typically, this is mixture model, in which mixture components represent known classes (64). Model-based approaches are computation-intensive and can be sensitive to assumptions made about the probability

model, but can provide a solid formal framework for the evaluation of many sources of uncertainty, and for assessing the probability of a sample belonging to a class.

3 Summary

A wide range of alternative approaches for clustering and classification of gene expression data are available. While differences in efficiency do exist, none of the well established approaches is uniformly superior to others. Choosing an approach requires consideration of the goals of the analysis, the background knowledge, and the specific experimental constraints. The quality of an algorithm is important, but is not in itself a guarantee of the quality of a specific data analysis. Uncertainty, sensitivity analysis and, in the case of classifiers, external validation or cross-validation should be used to support the legitimacy of results of microarray data analyses.

Acknowledgment

Work of Parmigiani partly supported by NSF grant NCI grant, 5P30 CA06973-39

References

[1] Speed TP, ed. *Statistical Analysis of Gene Expression Microarray Data*. London: Chapman and Hall 2003.

[2] Parmigiani G, Garrett ES, Irizarry RA, Zeger SL. The analysis of gene expression data: an overview of methods and software. In: Parmigiani G, Garrett ES, Irizarry RA, Zeger SL, eds., *The analysis of gene expression data: methods and software*, 1–45. New York: Springer 2003.

[3] Hartigan JA. *Clustering Algorithms*. Wiley 1975.

[4] Kaufmann L, Rousseeuw PJ. *Finding Groups in Data: An introduction to Cluster Analysis*. New york: Wiley 1990.

[5] Perou CM, Jeffrey SS, van de Rijn M, et al. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc Natl Acad Sci U S A* 96(16); 9212–9217 1999.

[6] Bullinger L, Dohner K, Bair E, et al. Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *New England Journal of Medicine* 350; 1605–1616 2004.

[7] Ripley BD. *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press 1996.

[8] Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer 2003.

[9] Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286; 531–537 1999.

[10] Ross DT, Scherf U, Eisen MB, et al. Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics* 24; 227–235 2000.

[11] Hedenfalk I, Duggan D, Chen Y, et al. Gene-expression profiles in hereditary breast cancer. *N Engl J Med* 344(8); 539–48 2001.

[12] Ihaka R, Gentleman R. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 5; 299–314 1996.

[13] Gentleman R. BioConductor: open source software for bioinformatics. `http://www.bioconductor.org` 2003.

[14] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Science USA* 95; 14863–14868 1998.

[15] Gordon AD. *Classification*. New York: Chapman and Hall/CRC 1999.

[16] Tamayo P, Slonim D, Mesirov J, et al. Interpreting gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Science USA* 96; 2907–2912 1999.

[17] Hartigan JA, Wong MA. A k-means clustering algorithm. *Applied Statistics* 28; 100–108 1979.

[18] Kohonen T. *Self-Organization and Associative Memory*. Berlin: Springer-Verlag 1989.

[19] Kachigan SK. *Multivariate Statistical Analysis: A Conceptual Introduction*. New York: Radius Press 1991.

[20] Dunteman GH. *Principal Components Analysis, Vol. 69*. Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-064. Newbury Park, CA: Sage 1989.

[21] Everitt B. *Applied Multivariate Data Analysis*. London: Edward Arnold 2001.

[22] Yeung KY, Ruzzo WL. Principal component analysis for clustering gene expression data. *Bioinformatics* 17; 763–774 2001.

[23] West M, Blanchette C, Dressman H, et al. Predicting the clinical status of human breast cancer using gene expression profiles. *Proceedings of the National Academy of Science USA* 98; 11462–11467 2001.

[24] Knudsen S. *A Biologist's Guide to Analysis of DNA Microarray Data*. New York: John Wiley and Sons 2002.

[25] Quackenbush J. Computational analysis of microarray data. *Nature Reviews Genetics* 2; 418–427 2001.

[26] Raychaudhuri S, Stuart JM, Altman RB. Principal components analysis to summarize microarray experiments: Application to sporulation time series. In: Altman RB, Dunker AK, Hunter L, Lauderdale K, Klein TE, eds., *Fifth Pacific Symposium on*

Biocomputing, 455–466 2000.

[27] Granucci F, Vizzardelli C, Pavelka N, et al. Inducible IL-2 production by dendritic cells revealed by global gene expression analysis. *Nature Immunology* 2; 882–888 2001.

[28] Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Science USA* 97(18); 10101–10106 2000.

[29] Khan J, Simon R, Bittner M, et al. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Research* 58; 5009–5013 1998.

[30] Bittner M, Meltzer P, Chen Y, et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 406; 536–540 2000.

[31] Yeung K, Fraley C, Murua A, Raftery A, Ruzzo W. Model-based clustering and data transformations for gene expression data. *Bioinformatics* 17; 977–987 2001.

[32] Kerr MK, Churchill GA. Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Science USA* 98; 8961–8965 2001.

[33] McShane LM, D RM, Freidlin B, Yu R, Li MC, Simon R. Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. Tech report #2, BRB, NCI, Bethesda, MD 2001.

[34] Bhattacharjee A, Richards WG, Staunton J, et al. Classification of human lung car-

cinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences USA* 98; 13790–13795 2001.

[35] National Research Council; Panel on Discriminant Analysis Classification and Clustering. *Discriminant Analysis and Clustering*. Washington, D. C.: National Academy Press 1988.

[36] Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. *JASA* 97; 77–87 2002.

[37] Gnanadesikan R. *Methods for Statistical Data Analysis of Multivariate Observations*. New York: Wiley 1977.

[38] Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group 1984.

[39] Hastie T, Tibshirani R. *Generalized Additive Models*. London: Chapman and Hall 1990.

[40] Neal RM. *Bayesian Learning for Neural Networks*. New York: Springer-Verlag 1996.

[41] Rios Insua D, Mueller P. Feedforward neural networks for nonparametric regression. In: *Practical Nonparametric and Semiparametric Bayesian Statistics*, 181–194. New York: Springer 1998.

[42] Slonim DK, Tamayo P, Mesirov P, Golub TR, Lander ES. Class prediction and discovery using gene expression data. Discussion paper, Whitehead/M.I.T. Center

for Genome Research, Cambrindge, MA 1999.

[43] Dettling M, Bühlmann P. Supervised clustering of genes. *Genome Biology* 3; 0069.1–0069.15 2002.

[44] Michie D, Spiegelhalter DJ, Taylor CC, eds. *Machine Learning, Neural and Statistical Classification*. New York: Ellis Horwood 1994.

[45] Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst* 95; 14–18 2003.

[46] DeGroot MH, Fienberg SE. The comparison and evaluation of forecasters. *The Statistician* 32; 12–22 1983.

[47] Toussaint GT. Bibliography on estimation of misclassification. *IEEE Transactions on Information Theory* IT-20; 472–79 1974.

[48] Radmacher MD, McShane LM, Simon R. A paradigm for class prediction using gene expression profiles. Tech report #1, BRB, NCI, Bethesda, MD 2001.

[49] Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Science USA* 99; 6567–6572 2002.

[50] Geman D, d'Avignon C, Naiman DQ, Winslow RL. Classifying gene expression profiles from pairwise mrna comparisons. *Statistical Applications in Genetics and Molecular Biology* 3; Article 19 2004.

[51] Cover TM, Hart PE. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* IT-13; 21–27 1967.

[52] Vapnik V. *Statistical Learning Theory*. New York: Wiley 1998.

[53] Burges CJC. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 2; 121–167 1998.

[54] Christianini N, Shawe-Taylor J. *An Introduction to Support-Vector Machines*. Cambridge: Cambridge University Press 2000.

[55] Lee Y, Lee CK. Classification of multiple cancer types by multicategory support vector machines using gene expression data. Tech. Rep. 1051, University of Wisconsin, Madison, WI 2002.

[56] Brown MPS, Grundy WN, Lin D, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Science USA* 97; 262–267 2000.

[57] Fisher RA. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7(part 2); 179–188 1936.

[58] Hastie TJ, Tibshirani R, Buja A. Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association* 89; 1255–1270 1994.

[59] Li W, Yang Y. How many genes are needed for a discriminant microarray data analysis? In: Lin SM, Johnson KF, eds., *Methods of Microarray Data Analysis*, 137–150. Dordrecht: Kluwer Academic 2002.

[60] Zhang H, Yu CY. Tree-based analysis of microarray data for classifying breast cancer. *Frontiers in Bioscience* 7; 63–67 2002.

[61] George EI, McCulloch RE. Variable selection via Gibbs sampling. *Journal of the American Statistical Association* 88; 881–889 1993.

[62] Clyde MA, DeSimone H, Parmigiani G. Prediction via orthogonalized model mixing. *Journal of the American Statistical Association* 91; 1197–1208 1996.

[63] Clyde MA, Parmigiani G. Bayesian variable selection and prediction with mixtures. *Journal of Biopharmaceutical Statistics* 8(3); 431–443 1998.

[64] Pavlidis P, Tang C, Noble WS. Classification of genes using probabilistic models of microarray expression profiles. In: Zaki MJ, Toivonen H, Wang JTL, eds., *Proceedings of BIOKDD 2001: Workshop on Data Mining in Bioinformatics*, 15–18. New York: Association for Computing Machinery 2001.

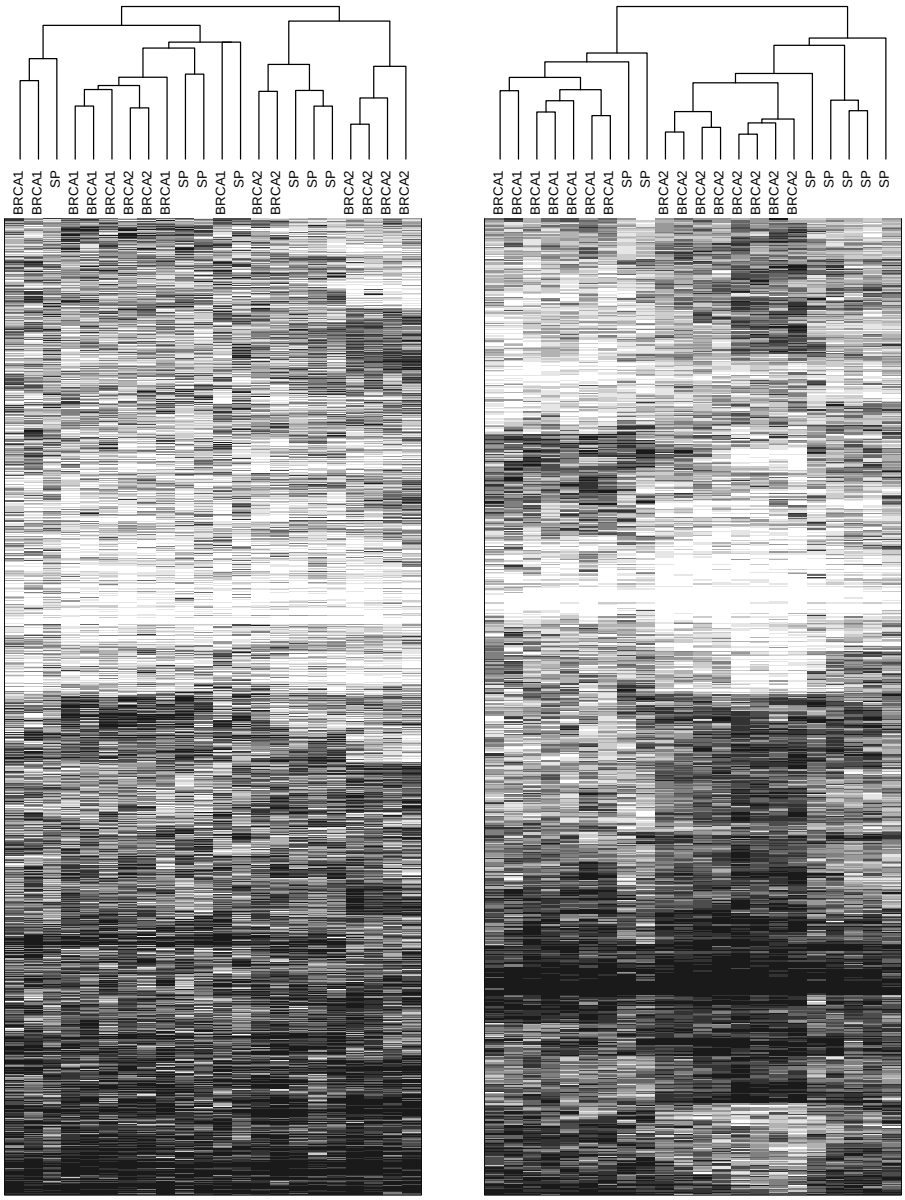

Figure 1: Hierarchical cluster analysis of the Hedenfalk breast cancer data. The grey scale image represents gene expression levels, with levels lower than the reference represented by white to light gray and levels higher than the reference represented by medium gray to black. The left panel includes all samples and genes. The right panel includes all samples and the top 25% genes most strongly associated with the presence of BRCA1 and BRCA2 mutations. The dendrograms for genes have been omitted.

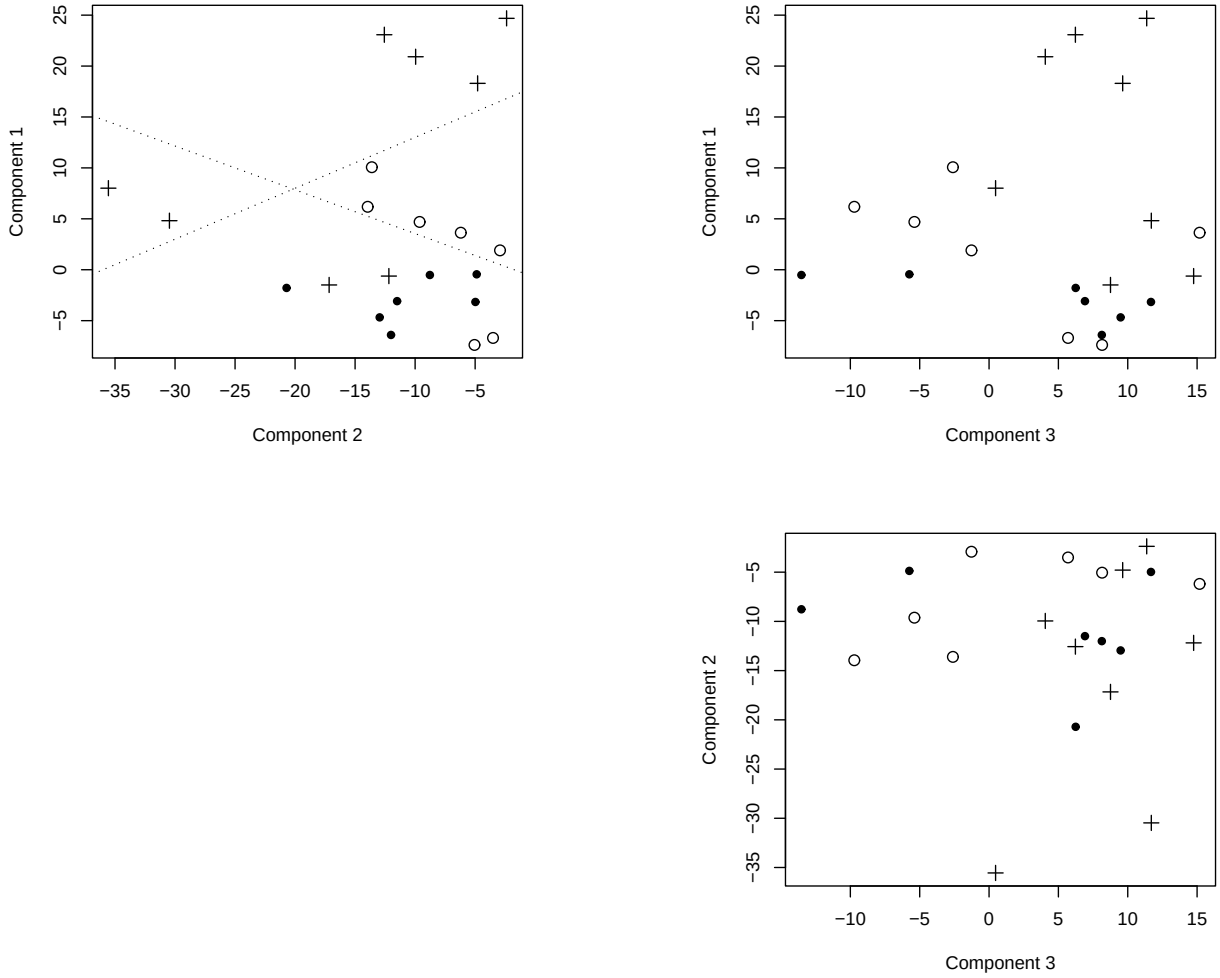

Figure 2: The first three principal components of the Hedenfalk breast cancer data. Open circles indicate sporadic samples, closed circles indicate BRCA1 samples, and plus symbols indicate BRCA2 samples. Dotted lines in the plot of component 1 versus component 2 distinguish the three types.

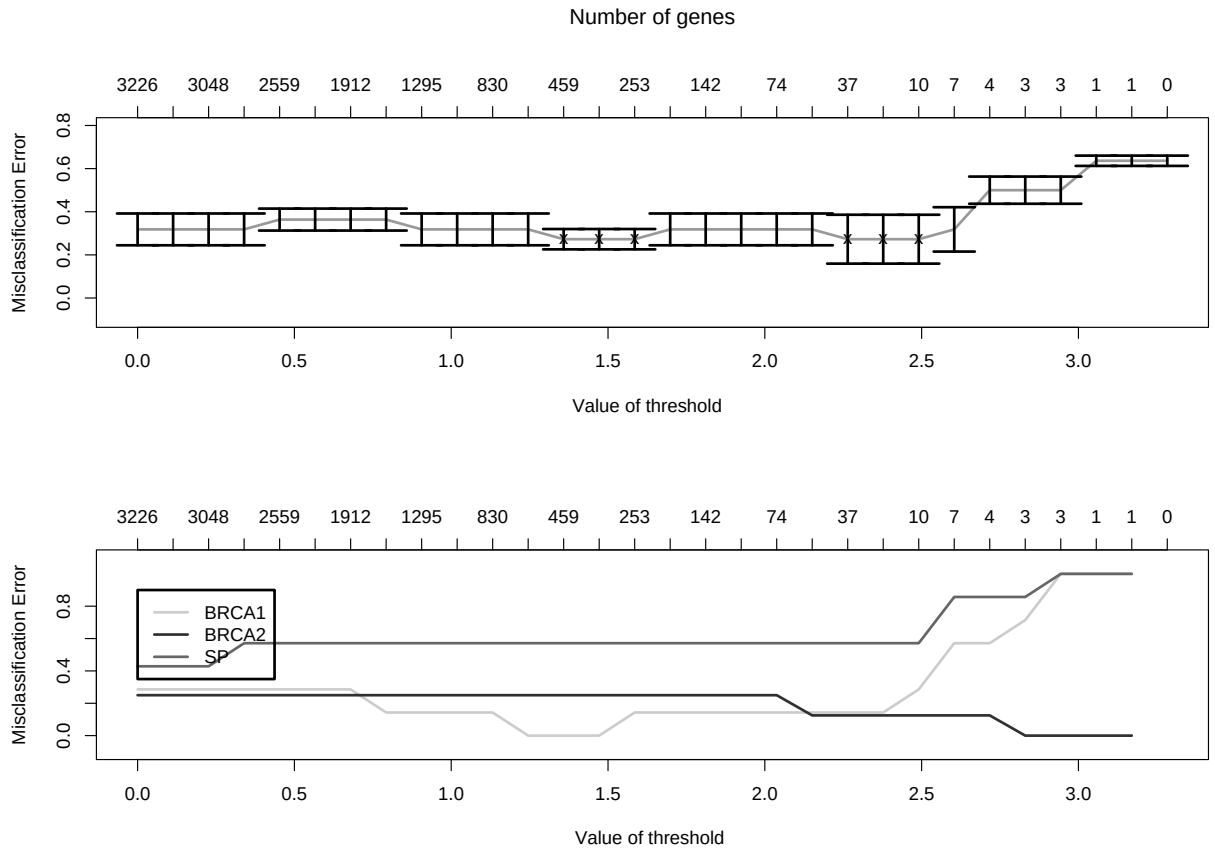

Figure 3: Misclassification error of PAM classifiers on the Hedenfalk breast cancer data. The top panel shows the overall classification error as a function of the threshold used to set centroid differences to zero. Classifiers on the right have a higher threshold and a more parsimonious use of genes. As the number of genes increases, the error rate decreases until about ten genes when the effects of overfitting offset the additional predictve ability of adding genes, and the error rates starts increasing. The bottom panel shows the classification error by class.