

# Supplementary material to “Molecular Signatures from Gene Expression Data”

Ramón Díaz-Uriarte  
Statistical Computing Team  
Structural Biology and Biocomputing Programme  
Spanish National Cancer Center (CNIO)  
Melchor Fernández Almagro 3  
Madrid, 28029  
Spain.  
rdiaz02@gmail.com  
<http://ligarto.org/rdiaz>

**Running Head:** Gene expression signatures.

# 1 A review of previous methods for identifying molecular signatures

Here we provide a quick review of some of the most relevant methods for finding signatures, emphasizing how they fail to satisfy some of the signature requirements. Some of the papers included here do not set as their objectives the identification of molecular signatures, but because of the methods used, or the contexts in which they have been cited, or their similarity with methods that are used to find molecular signatures, it seemed appropriate to include them in this section.

## 1.1 Methods that return a single signature component or several components without inherent meaning

In these cases, the component is a weighted average of the expression of a set of genes. The main drawback of these methods is that components do not necessarily show tight coexpression. [1] proposed the **compound covariate** method; they form a predictor by weighting pre-selected genes using t-statistics; a problem of this implementation is that extension to problems other than classification into two groups is not straightforward. The compound covariate method has been applied by [2, 3]. A similar approach uses **Diagonal linear discriminant analysis (DLDA)**, a simple type of discriminant analysis, that has been shown to perform remarkably well [4, 5]; DLDA and compound covariate, in the two class problem, yield very similar results (there are small differences in the weights of genes). DLDA is restricted to classification problems (without limit on the number of classes). Another popular approach is the **weighted gene voting** of [6], used also, for instance, by [7, 8]. This method returns a component that is the weighted average of the expression of a set of pre-selected genes; the weights are given by Golub et al.’s “signal-to-ratio” expression (which resembles the DLDA expression, except they use an unusual estimate of the standard error of the difference in expression —see [4]). Extension of this method to problems other than classification into two groups is not straightforward. [9] developed the **shrunk centroids** method, which is also closely related to the three previous methods; their approach is to find a “shrunk centroid” for each class (i.e., the coefficients or weights of most genes are shrunk towards 0) and then classify a sample to its closest centroid. It is not clear how this method could be used in problems other than classification. Somewhat similar to shrunk centroids, but worded and obtained in a very different way, is the method of [10] and [11]: an individual is assigned to the group with which its gene expression profile is most highly correlated; the profile of a group is formed by the set of genes that shows strong correlation with outcome (like a t-test). Extension to problems other than 2-class classification is unclear. Many of the methods above use some form of feature selection; [12] proposed that, instead of using statistics for each gene on its own, **gene pair ranking** be used.

## 1.2 Several signature components: averages

Rosenwald et al. [13] obtain each component as a simple average of “significant genes” within a signature. Signatures had been previously defined, via clustering, in [14]. Their method was originally devised for survival analysis (Cox model) but could be used for classification. A similar approach was used by Baechler et al. [15], where genes that belong to a signature were found using gene expression fold change and absolute difference with respect to controls. Using simple averages discards information about the relative prediction strength of different genes; more importantly, the signatures or signature components need to have been previously defined.

## 1.3 Penalized likelihood and ridge regression-like

Eilers et al. [16] use this approach for classification with logistic regression and Pawitan et al. [17] use it for survival analysis using Cox regression. Basically, these authors build a model that can include all genes, but drive many coefficients close to 0 using penalization (these methods have connections with ridge regression and random effects models). For computational reasons both use SVD before model fitting. The problem of SVD (and PCA) on the complete set of genes is that it precludes simple interpretation because all genes have loadings on every component. SVD, however, is not necessary, as shown in [18] with several microarray data sets. Regardless of SVD, these methods can return coefficients of the original genes, but there is no sense in which they return components of genes with tight coexpression. Dettling and Bühlman [19] combine penalized logistic regression with gene clustering, without using SVD or PCA, and use as predictors gene cluster centroids; however, the optimized criteria in their model fitting does not include tight gene coexpression.

## 1.4 Partial Least Squares

Partial Least Squares (PLS) resemble PCA, but components are obtained so that the covariance with the dependent variable is maximized [20–22]. Ghosh [23], Nguyen and Rocke [24], and Gusnanto et al. [18] use PLS for classification; Ghosh uses optimal scoring to deal with the classification problem, whereas Nguyen and Rocke and Gusnanto et al. first run PLS on the recoded variable (i.e., dummies for the classes), and then use a standard classifier (e.g., logistic discrimination). The use of PLS for survival analysis with microarray data is examined by Park et al. [25]. The PLS solution forces each of the linear combinations of original variables to have a sample correlation of 0, which is likely to be an inappropriate requirement for molecular signatures; more importantly, the main problem of PLS is similar to that of PCA on the complete set of genes: difficulty in the interpretation of components because each component has loadings of all genes.

## 1.5 Sufficient dimension reduction: SIR, SAVE, MAVE

Sufficient dimension reduction methods attempt to find a linear combination of the genes that constitute a “sufficient subspace” (in the sense that the conditional distribution of  $Y$ , the dependent variable, given the sufficient subspace, is independent of the original predictors). SIR, SAVE and MAVE are three methods for finding the sufficient subspace. This family of methods is appealing because it does not require us to specify the functional relationship between dependent and independent variables. The main problem of these methods is, again, the interpretation of the resulting components, since all genes have “loadings” in every component. Chiaromonte and Martinelli [26] apply SIR to microarray data; Bura and Pfeiffer [27] apply SIR and SAVE, and Antoniadis et al. [28] use MAVE. Chiaromonte & Martinelli and Bura & Pfeiffer need to go through a previous SVD before applying SIR or SAVE because, otherwise, they cannot invert  $\hat{\Sigma}$ ; Antoniadis et al. do not use SVD but, for computational reasons, they need to preselect a subset of genes.

## 1.6 Bayesian methods

The group of M. West and collaborators have used bayesian methods for finding supergenes. In [29, 30] they first apply SVD and then use probit regression with bayesian regularization. In [31] they first use SVD on clusters of genes and then apply bayesian ensembles of classification trees. As with other methods that use SVD or PCA, interpretation is complicated and metagenes are not made of only tightly coexpressed genes; in addition, for the second method, there is some arbitrariness in the selection of the number of clusters and ensembles of trees are difficult to interpret. Finally, both methods might have limited use because of the difficulties of selecting priors and setting up the Markov Chain Monte Carlo computation. A somewhat related approach is that of [32], although they make no attempt to obtain closely related sets of genes, but concentrate on model selection for classification.

## 1.7 Other methods

### 1.7.1 Support Vector Machines

Support vector machines have been applied to microarray data [33–36]. Although SVM are excellent classifiers [37], they are difficult to interpret and, more important in our context, there is nothing like a signature component ever returned (and thus no sense in which tight gene coexpression is attempted).

### 1.7.2 Supervised harvesting of expression trees

This method was proposed by Hastie et al. [38]. The authors obtain all  $2p - 1$  clusters of genes using hierarchical clustering. Then, using a forward addition-like method they build a model of size  $M$  (possibly including interactions). Finally, they prune the model down, and select model size using cross-validation. This approach can be applied to any type of dependent variable (be it in a classification, regression, or

survival analysis problem). Supervised harvesting shows similarities with the method suggested here. However, with supervised harvesting of expression trees, there is no need for clusters of genes to be of tightly coexpressed genes, and clusters are formed without using information from the dependent variable.

### 1.7.3 Wilma: supervised clustering

This method has been developed by Dettling and Bühlman [39]. The information from the dependent variable is explicitly included in the formation of clusters of genes. Each cluster of genes is a simple linear combination of genes (i.e., weights are 1 or -1), and genes are added to a cluster until predictive performance (evaluated with two different statistics) decreases. The number of clusters of genes is to be specified by the user (i.e., it is not an outcome of the algorithm), and this algorithm would be difficult to extend to problems other than class prediction. More importantly, tight coexpression of genes within a cluster is not an explicit objective of the algorithm: because of the way genes are combined, this algorithm tends to return clusters of genes that show coexpression, but how tightly coexpressed genes are in each metagene cannot be specified in advance.

### 1.7.4 Block PCA

This method has been developed by Liu et al. [40]. They divide genes in blocks according to correlation (they use cluster analysis) so as to obtain blocks of high correlation among the genes in that block. Within each block they perform PCA and select “important” genes. Next, they use a second PCA with the selected genes from the previous step. Finally, they use these components for classification. Thus, Liu et al. [40] explicitly try to obtain blocks with high internal correlation. However, the clusters of genes —both number and composition— are formed without using information from the dependent variable; in addition, their results show multiple components being selected from each block (as many as 11 to 16), which makes it very difficult to interpret results in terms of a few sets of highly correlated genes.

Somewhat related to this approach are the different methods that attempt to obtain “simple” PCAs, such as in [41–43], and references therein. If each component of these simple PCAs has loadings of only a few genes, then interpretation will be greatly enhanced. However, the search for this simple PCAs is carried out without any information from the dependent variable (i.e., the objective is to understand the structure of the independent variables), and thus might not be appropriate when good predictive performance is required.

### 1.7.5 Rank-ordered PCA with gene selection

Landgrebe et al. [44] build upon an idea of Krzanowski [45]. They use “rank-ordered PCA”, a PCA followed by an ordering of the components as a function of their relevance for separating groups. Next, only relevant components are retained, and the

“most important” genes from those components are selected, and used for classification. Interestingly, [46] published another approach to PCA with class structure that seems to supersede his 1992 method; it seems that using Krzanowski’s 1995 method would only require modification of the first step of the algorithm in Landgrebe et al. More importantly, the genes returned do not constitute a set of tightly coexpressed genes (and it is difficult to extend this method to problems other than classification).

## 2 Extensions to the algorithm

### 2.1 Other classifiers and other types of dependent and independent variables

Including other classifiers is simple: we only need to use the appropriate method in steps 1 and 5b of the algorithm. In addition, it is straightforward to select seed genes or exclude genes from signature using other measures of predictive performance, such as Brier’s score or Sommers’  $D_{xy}$ , instead of misclassification rates (see [53, p. 248] for reasons why one might not want to use misclassification error rates). The use of alternatives measures of predictive performance can be particularly interesting when many genes yield an error of 0; in these cases, when using DLDA, we might want, instead, to use separation of groups, and not just prediction error, in particular when searching for variables that are best for describing differences among the groups [54, p. 390]; measures such as Hotelling’s and Wilks (on, specially, the out-of-sample points) seem appropriate.

However, use of other classifiers should probably try to preserve simplicity of interpretation; in this context, then, support vector machines might not be a good choice, whereas classification trees or other types of discriminant analysis might be a interesting alternative.

Extension to other types of dependent variables requires only deciding on the appropriate criterion of prediction error rate. With multiple regression an immediate choice could be prediction residual standard error. The choice is more complicated for survival analysis (see [53] for alternatives). In addition, fitting survival models is computationally more costly than DLDA or NN; thus, instead of using a true cross-validatory procedure to decide on seed genes or genes to remove from a signature, a statistic such as AIC (which does not require repeated fitting, but which will asymptotically lead to the same model selection as cross-validation —e.g. [55]) might be preferred. We are currently working in this area. Finally, forcing the use of other independent variables (e.g., sex or age of patients) in the models is straightforward. An open issue is whether “de-noising” of gene expression values prior to analysis, such as can be achieved using the “probability of expression” model of Parmigiani and collaborators [56,57], can be of help for the performance of the signature method.

## 2.2 Identifying outlying and influential points

Once the scores of each individual on each signature have been obtained, we can apply the usual diagnostics for the statistical method where the scores are used as the predictors; therefore, these diagnostics are obtained taking the signature scores as given. However, it might be more interesting to identify influential and outlying points in the process of signature component definition. Jolliffe [43] discusses outlier and influential points in PCA. Of particular relevance for us, with regards to influential points, is the change in the genes that compose each signature. Outlier and influential measures that are meaningful in the context of signatures are relevant because we thus obtain an indication of samples that affect our very perception of signatures and sets of coexpressed genes. Further work on this area is in progress.

## 3 Results: can we recover signatures when they are present?

The results for DLDA and NN as the underlying predictors are shown in figures 1 and 2.

## 4 Results: comparison with other classification methods on real microarray data sets

The results change very little from using DLDA vs. NN, or setting  $c_1 = c_2 = 1$  vs.  $c_1 = c_2 = 0$ . We show figures for four possible combinations in Figures 3 and 4.

## 5 Comparing orientation of within-group subspaces using Krzanowski's approach

With our data sets, only for the NCI 60 and the Leukemia data sets we obtain relatively stable signature components involving a similar set of genes. If we compare the first signature component between the two groups of patients of the Leukemia data set, we find an angle of  $47^\circ$  between subspaces. As originally proposed, this is a descriptive tool because there is no formal definition of what a large angle is; although  $47^\circ$  seems quite a large deviation between subspaces, it is unclear if it is really such a large deviation if we take into account the small sample sizes (27 and 11 subjects per group).

This problem can be much better illustrated with the NCI 60 data set, because the first signature component is made up of only two genes, and thus we can represent the data using a scatterplot. When there are more than two groups of subjects, Krzanowski [52] shows how we can carry out a simultaneous comparison of the principal components obtained in all the groups. However, to allow for easier

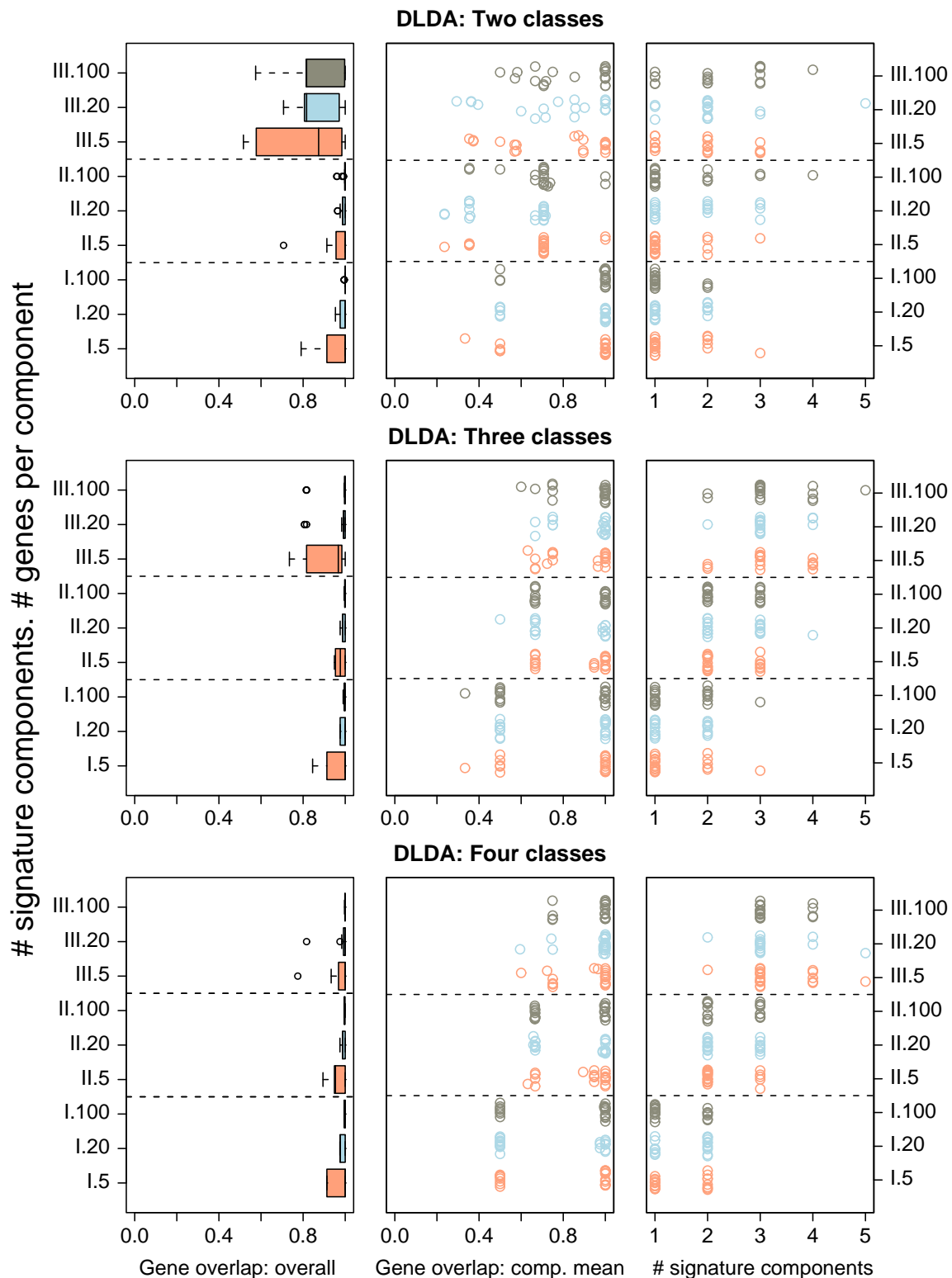


Figure 1: Signature recovery when using the signature method with the DLDA predictor. Each line represents a combination of number of signatures and number of genes per signature. For instance, II.20 denotes II signature components with 20 genes per signature component. Based on 20 replicate simulations. To facilitate distinguishing data, points have been jittered vertically in the center and right panels. See text for explanation of variables.

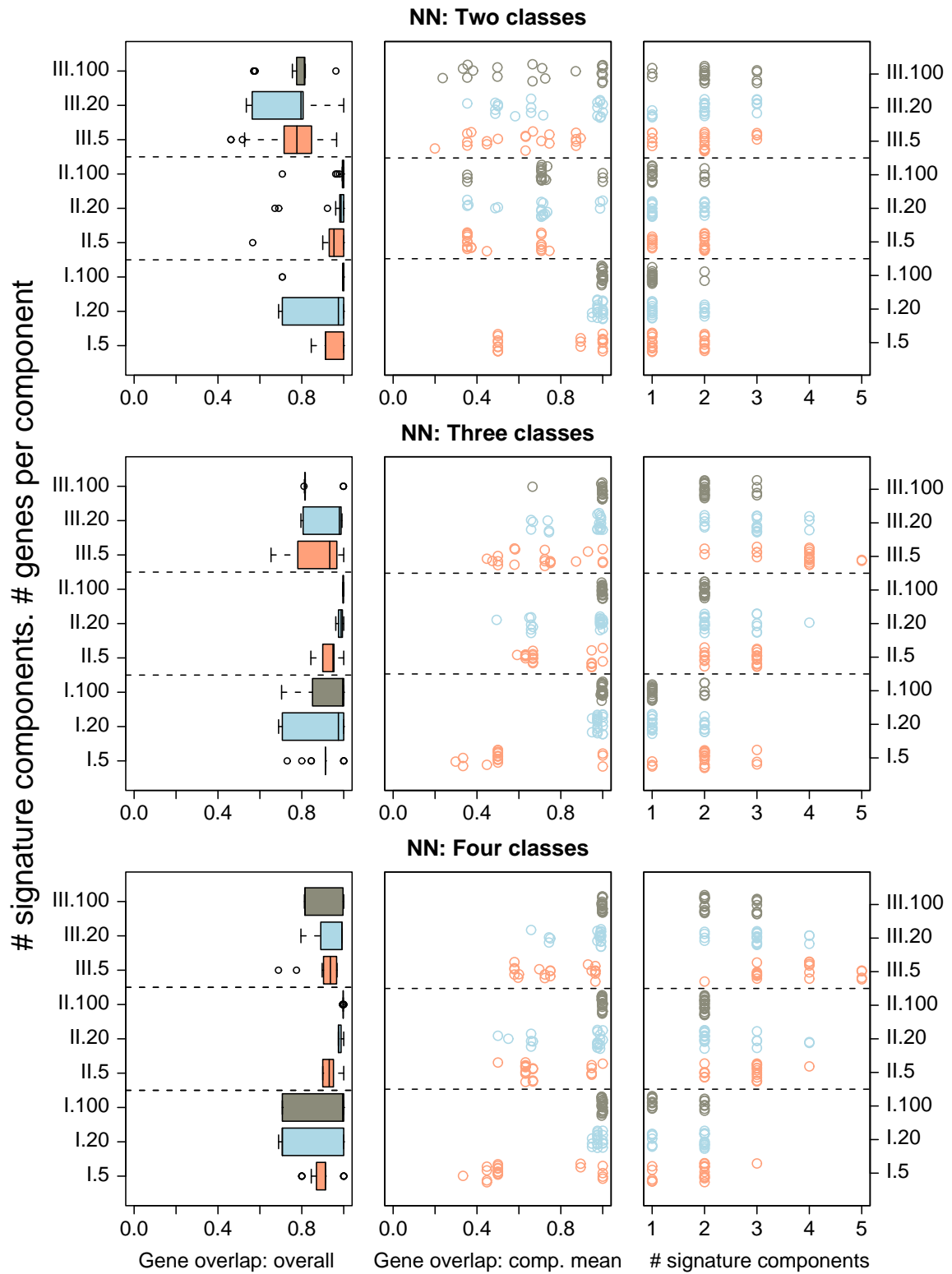


Figure 2: Signature recovery when using the signature method with the NN predictor. See figure 1 and text for explanation of variables.

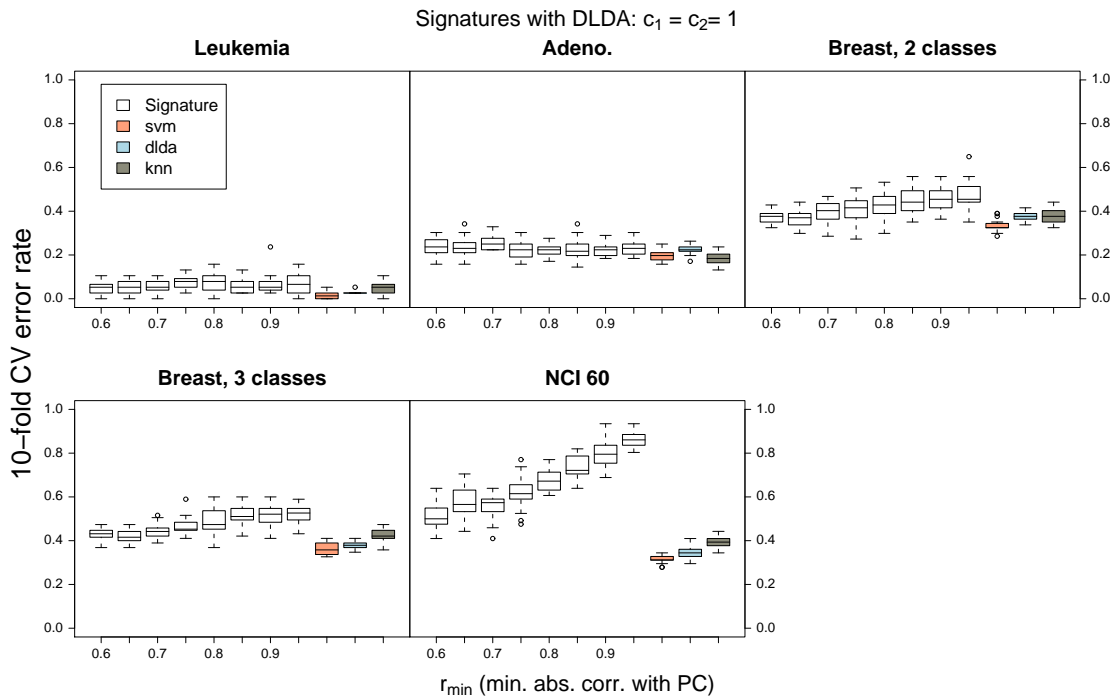
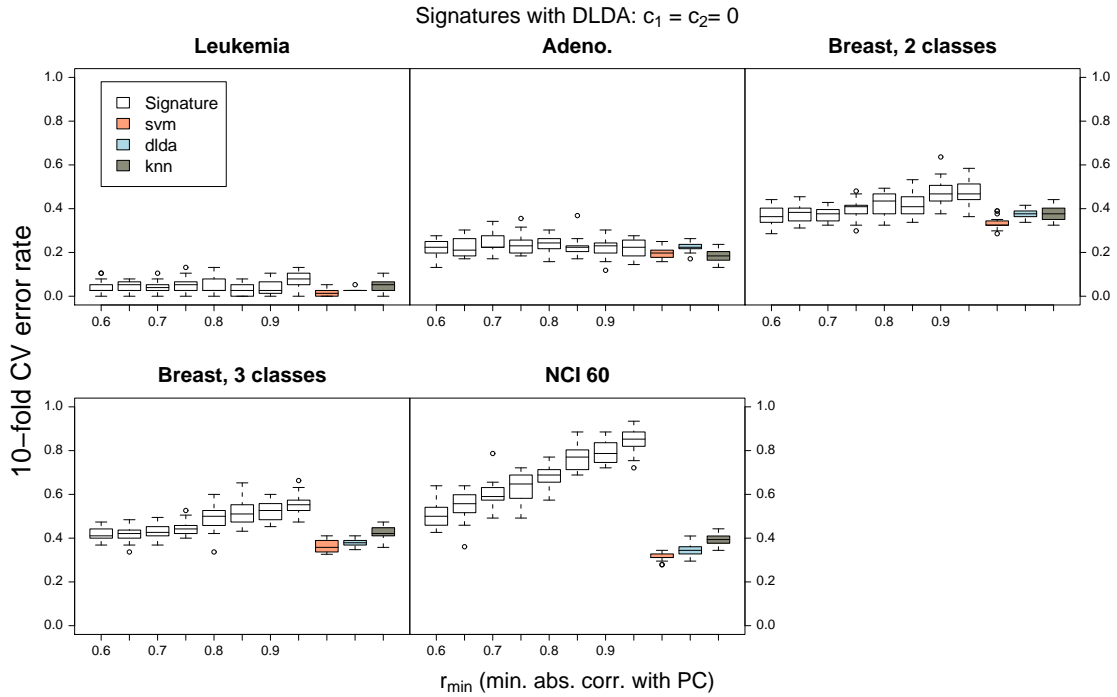


Figure 3: Predictive performance, as a function of  $r_{min}$ , of the signature method using DLDA as classifier and comparison with SVM, KNN, and DLDA. Figures based on 20 replicates of the 10-fold-CV procedure.

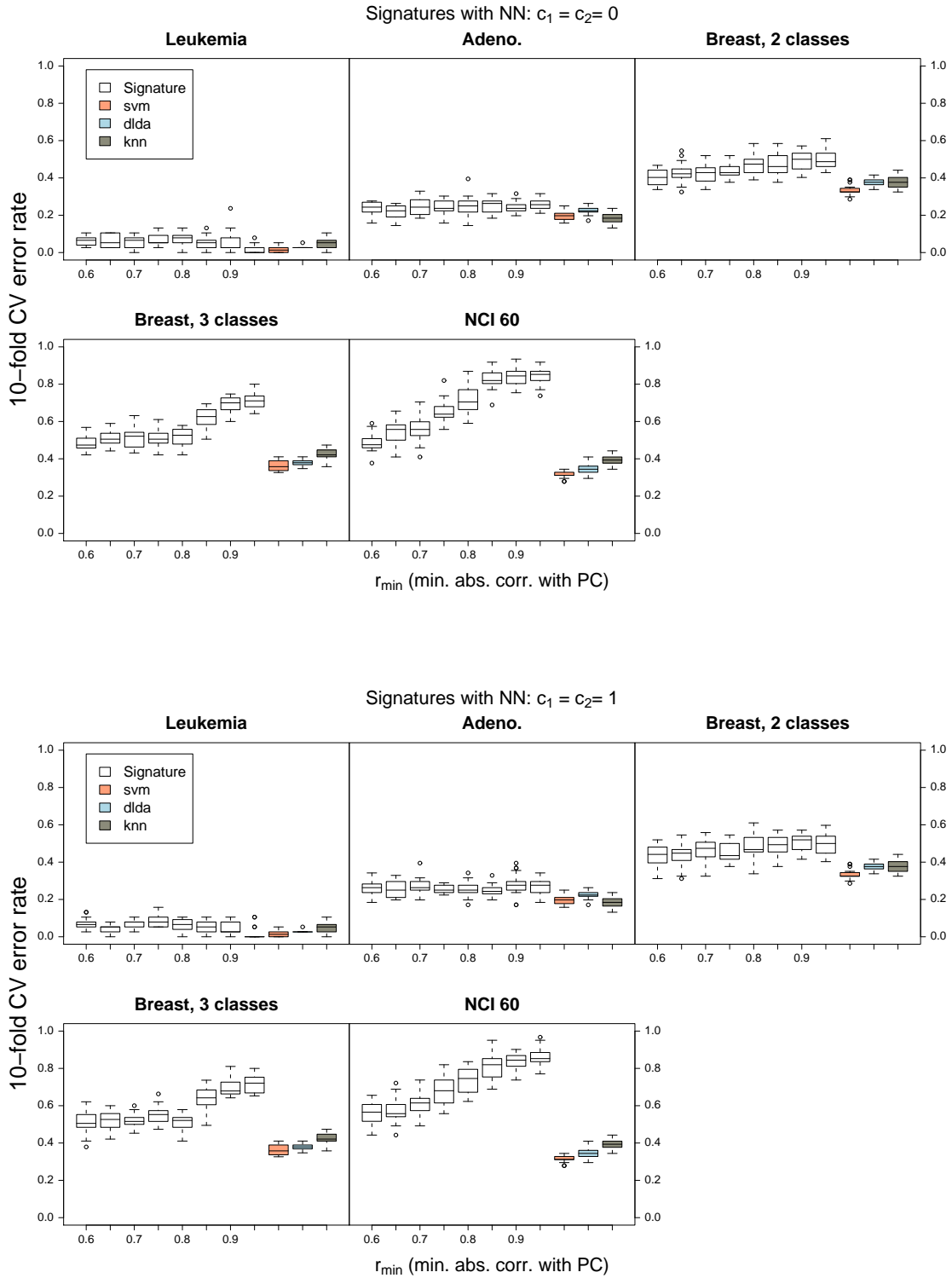


Figure 4: Predictive performance, as a function of  $r_{min}$ , of the signature method using NN as classifier and comparison with SVM, KNN, and DLDA. Figures based on 20 replicates of the 10-fold-CV procedure.

visualization, we will carry out all the possible two-group comparisons. The data are shown in Figure 5, using different colors to represent different groups; the triangular matrix inserted shows the angles between subspaces of every pair of groups. Several features are to be mentioned; first, there seems to be a clear relationship between the two genes; second, the angles between subspaces for some groups are very small (dark green, dark blue, red, purple), and seem to follow along the same direction as the pooled sample; third, however, the angles are very large between some other groups (e.g., light blue with red or purple); fourth, we can wonder how much we can trust these angle comparisons with such small sample sizes.

### NCI 60, first signature component

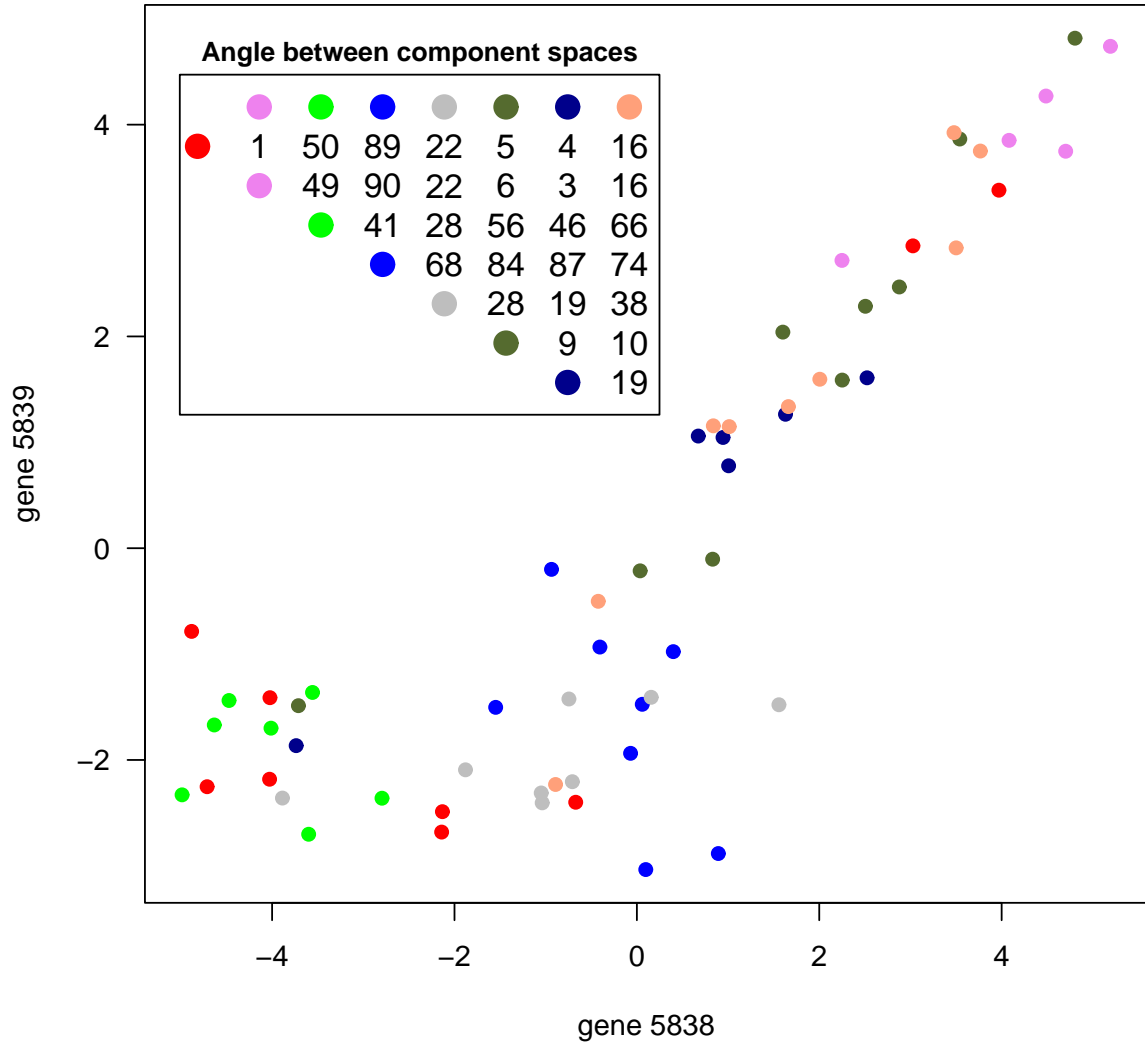


Figure 5: Scatterplot of the first signature component and angle between component spaces. The scatterplot shows the values of the genes that form the first signature component for the NCI 60 data set, with circles of different colors indicating the eight different groups of patients. The inserted table shows the angles (rounded to the nearest integer) between every pair of component spaces (i.e., between the component spaces for each group of patients) using the first eigenvector; for example, the angle between the component spaces of subjects in the “purple group” and subjects in the “gray group” is 22 degrees. See text for details.

## References

- [1] Radmacher MD, McShane LM, Simon R: **A paradigm for class prediction using gene expression profiles.** *J Comp Biol* 2002, 9:505–511.
- [2] Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi O, Wilfond B, Borg A, Trent J: **Gene-expression profiles in hereditary breast cancer.** *N Engl J Med* 2001, 344:539–548.
- [3] Wright G, Tan B, Rosenwald A, Hurt EH, Wiestner A, Staudt LM: **A gene expression-based method to diagnose clinically distinct subgroups of diffuse large b cell lymphoma.** *Proc Natl Acad Sci USA* 2003, 100:9991–9996.
- [4] Dudoit S, Fridlyand J, Speed TP: **Comparison of discrimination methods for the classification of tumors using gene expression data.** *J Am Stat Assoc* 2002, 97:77–87.
- [5] Romualdi C, Campanaro S, Campagna D, Celegato B, Cannata N, Toppo S, Valle G, Lanfranchi G: **Pattern recognition in gene expression profiling using dna array: a comparative study of different statistical methods applied to cancer classification.** *Hum Mol Genet* 2003, 12:823–836.
- [6] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, 286:531–537.
- [7] Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RCT, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, Ray TS, Koval MA, Last KW, Norton A, Lister TA, Mesirov J, Neuberg DS, Lander ES, Aster JC, Golub TR: **Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning.** *Nature Medicine* 2002, 8:68–74.
- [8] Ramaswamy S, Ross KN, Lander ES, Golub TR: **A molecular signature of metastasis in primary solid tumors.** *Nature Genetics* 2003, 33:49–54.
- [9] Tibshirani R, Hastie T, Narasimhan B, Chu G: **Diagnosis of multiple cancer types by shrunken centroids of gene expression.** *Proc Natl Acad Sci USA* 2002, 99:6567–6572.
- [10] van Belle G: **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature* 2002, 415:530–536.
- [11] van de Vijver MJ, He YD, van 't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S,

- Rutgers ET, Friend SH, Bernards R: **A gene-expression signature as a predictor of survival in breast cancer.** *N Engl J Med* 2002, 347:1999–2009.
- [12] Bø TH, Jonassen I: **New feature subset selection procedures for classification of expression profiles.** *Genome Biology* 2002, 3:0017.1–0017.11.
- [13] Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, Gascoyne RD, Muller-Hermelink HK, Smeland EB, Giltane JM, Hurt EM, Zhao H, Averett L, Yang L, Wilson WH, Jaffe ES, Simon R, Klausner RD, Powell J, Duffey PL, Longo DL, Greiner TC, Weisenburger DD, Sanger WG, Dave BJ, Lynch JC, Vose J, Armitage JO, Montserrat E, Lopez-Guillermo A, Grogan TM, Miller TP, LeBlanc M, Ott G, Kvaloy S, Delabie J, Holte H, Krajci P, Stokke T, Staudt LM, the Lymphoma/Leukemia Molecular Profiling Project: **The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma.** *N Engl J Med* 2002, 346:1937–1947.
- [14] Shaffer A, Rosenwald A, Hurt E, Giltane J, Lam L, Pickeral O, Staudt L: **Signatures of the immune response.** *Immunity* 2001, 15:375–385.
- [15] Baechler EC, Batliwalla FM, Karypis G, Gaffney PM, Ortmann WA, Espe KJ, Shark KB, Grande WJ, Hughes KM, Kapur V, Gregersen PK, Behrens TW: **Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus.** *Proc Natl Acad Sci USA* 2003, 100:2610–2615.
- [16] Eilers PHC, Boer JM, van Ommen GJ, van Houwelingen HC: **Classification of microarray data with penalized logistic regression.** *Proceedings of SPIE volume 4266: progress in biomedical optics and imaging* 2001, San José.
- [17] Pawitan Y, Bjöhle J, Wedren S, Humphreys K, Skoog L, Huang F, Amler L, Shaw P, Hall P, Bergh J: **Gene expression profiling for prognosis using cox regression.** *Statist Med* 2004, 23:1767–1780.
- [18] Gusnanto A, Pawitan Y, Ploner A: **Variable selection in gene and protein expression data.** Technical report, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, 2003.
- [19] Dettling M, Bühlmann P: **Finding predictive gene groups from microarray data.** *J Multivariate Anal* 2004, 90:106–131.
- [20] Garthwaite PH: **An interpretation of partial least squares.** *J Am Stat Assoc* 1994, 89:122–127.
- [21] Stone M, Brooks RJ: **Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression (with discussion).** *J R Stat Soc B* 1990, 52:237–269.

- [22] Frank IE, Friedman JH: **A statistical view of some chemometrics regression tools.** *Technometrics* 1993, 35:109–135.
- [23] Ghosh D: **Penalized discriminant methods for the classification of tumors from gene expression data.** *Biometrics* 2003, In press.
- [24] Nguyen DV, Rocke DM: **Multi-class cancer classification via partial least squares with gene expression profiles.** *Bioinformatics* 2002, 18:1216–1226.
- [25] Park PJ, Tian L, Kohane IS: **Linking gene expression data with patient survival times using partial least squares.** *Bioinformatics* 2002, 18, S1:S120–S127.
- [26] Chiaromonte F, Martinelli J: **Dimension reduction strategies for analyzing global gene expression data with a response.** *Mathematical Biosciences* 2002, 176:123–144.
- [27] Bura E, Pfeiffer RM: **Graphical methods for class prediction using dimension reduction techniques on dna microarray data.** *Bioinformatics* 2003, 19:1252–1258.
- [28] Antoniadis A, Lambert-Lacroix S, Leblanc F: **Effective dimension reduction methods for tumor classification using gene expression data.** *Bioinformatics* 2003, 19:563–570.
- [29] West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JAJ, Marks JR, Nevins JR: **Predicting the clinical status of human breast cancer by using gene expression profiles.** *Proc Natl Acad Sci USA* 2001, 98:11462–11467.
- [30] Huang E, Ishida S, Pittman J, Dressman H, Bild A, Kloos M, D’Amico M, Pestell RG, West M, Nevins JR: **Gene expression phenotypic models that predict the activity of oncogenic pathways.** *Nature Genetics* 2003, 34:226–230.
- [31] Huang E, Cheng SH, Dressman H, Pittman J, Tsou MH, Horng CF, Bild A, Iversen ES, Liao M, Chen CM, West M, Nevins JR, Huang AT: **Gene expression predictors of breast cancer outcomes.** *Lancet* 2003, 361:1590–1596.
- [32] Sha N, Vannucci M, Brown PJ, Trower MK, Amphlett G, Falciani F: **Gene selection in arthritis classification with large-scale microarray expression profiles.** *Comp Funct Genom* 2003, 4:171–181.
- [33] Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D: **Support vector machine classification and validation of cancer tissue samples using microarray expression data.** *Bioinformatics* 2000, 16:906–914.

- [34] Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang C, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov J, Poggio T, Gerald W, Loda M, Lander E, Golub T: **Multiclass cancer diagnosis using tumor gene expression signatures.** *Proc Natl Acad Sci USA* 2001, 98:15149–15154.
- [35] Lee Y, Lee CK: **Classification of multiple cancer types by multicategory support vector machines using gene expression data.** *Bioinformatics* 2003, 19:1132–1139.
- [36] Su A, Welsh J, Sapinoso L, Kern S, Dimitrov P, Lapp H, Schultz P, Powell S, Moskaluk C, Frierson HJ, Hampton G: **Molecular classification of human carcinomas by use of gene expression signatures.** *Cancer Research* 2001, 61:7388–7393.
- [37] Hastie T, Tibshirani R, Friedman J: *The elements of statistical learning.* New York: Springer, 2001.
- [38] Hastie T, Tibshirani R, Botstein D, Brown P: **Supervised harvesting of expression trees.** *Genome Biology* 2001, 2:0003.1–0003.12.
- [39] Dettling M, Bühlmann P: **Supervised clustering of genes.** *Genome Biology* 2002, 3:0069.1–0069.15.
- [40] Liu A, Zhang Y, Gehan E, Clarke R: **Block principal component analysis with application to gene microarray data classification.** *Statist Med* 2002, 21:3465–3474.
- [41] Rousson V, Gasser T: **Simple component analysis.** Technical report, Department of Biostatistics, University of Zürich, Switzerland, 2003.
- [42] Vines SK: **Simple principal components.** *Applied Statistics* 2000, 49:441–451.
- [43] Jolliffe IT: *Principal component analysis, 2nd ed..* New York: Springer, 2002.
- [44] Landgrebe J, Wurst W, Welzl G: **Permutation-validated principal components analysis of microarray data.** *Genome Biology* 2002, 3:0019.1–0019.11.
- [45] Krzanowski WJ: **Ranking principal components to reflect group structure.** *J Chemometrics* 1992, 6:97–102.
- [46] Krzanowski WJ: **Orthogonal canonical variates for discrimination and classification.** *J Chemometrics* 1995, 9:509–520.
- [47] Braga-Neto U, Hashimoto R, Dougherty ER, Nguyen DV, Carroll RJ: , .
- [48] Ripley BD: *Pattern recognition and neural networks.* Cambridge: Cambridge University Press, 1996.

- [49] Lim TS, Loh WY, Shih YS: **A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms.** *Machine Learning* 2000, 40:203–228.
- [50] Ambroise C, McLachlan GJ: **Selection bias in gene extraction on the basis of microarray gene-expression data.** *Proc Natl Acad Sci USA* 2002, 99:6562–6566.
- [51] Morrison DF: *Multivariate statistical methods*. New York: McGraw-Hill, 1990.
- [52] Krzanowski WJ: *Principles of multivariate analysis*. Oxford: Oxford University Press, 1998.
- [53] Harrell JFE: *Regression modeling strategies*. New York: Springer, 2001.
- [54] McLachlan GJ: *Discriminant analysis and statistical pattern recognition*. New York: Wiley, 1992.
- [55] Pawitan Y: *In all likelihood: statistical modelling and inference using likelihood*. Oxford: Oxford University Press, 2001.
- [56] Parmigiani G, Garrett E, Anbazhagan R, Gabrielson E: **A statistical framework for expression-based molecular classification in cancer.** *J Royal Statistical Society, Series B* 2002, 64:717–736.
- [57] Scharpf R, Garrett E, Hu J, Parmigiani G: **Statistical modeling and visualization of molecular profiles in cancer.** *BioTechniques* 2003, 34:S22–S29.