Biplots in Practice

MICHAEL GREENACRE

Professor of Statistics at the Pompeu Fabra University

Chapter 13 Offprint

CASE STUDY–BIOMEDICINE

Comparing Cancer Types According to Gene Expression Arrays

First published: September 2010
ISBN: 978-84-923846-8-6

Supporting websites:
http://www.fbbva.es
http://www.multivariatestatistics.org

CHAPTER 13

Case Study 1: Comparing Cancer Types According to Gene Expression Arrays

This first case study contains many aspects of biplots treated in this book. The context is a large data set of microarray data from tumour samples found in children. This is a very "wide" data set in the sense that there are only 63 samples but over 2000 variables in the form of genes expressed in the microarray experiments. The variables are on the same continuous scale and so the regular PCA biplot of Chapter 6 will be used to visualize the raw data. But because the samples are grouped we shall also apply the centroid biplot described in Chapter 11 to show separation of the tumour groups. There are two additional aspects to this case study. First, because of the large number of variables we will be interested in quantifying the contributions of each one to the biplots that we construct, with a view to reducing the gene set to the most important ones. Second, an additional sample of 20 tumours is available, which can be used to test whether the biplot provides a prediction rule capable of classifying these additional tumours correctly.

Contents

Data set "cancer"

This data set "cancer" is taken from the book *The Elements of Statistical Learning (second edition)* by Hastie, Tibshirani and Friedman and consists of a matrix of 2308 genes (columns) observed on 63 samples (rows)—see the Bibliography for a link to the book's website and accompanying data sets. The data arise from microarray experiments, a technology which has become important in genomic research,

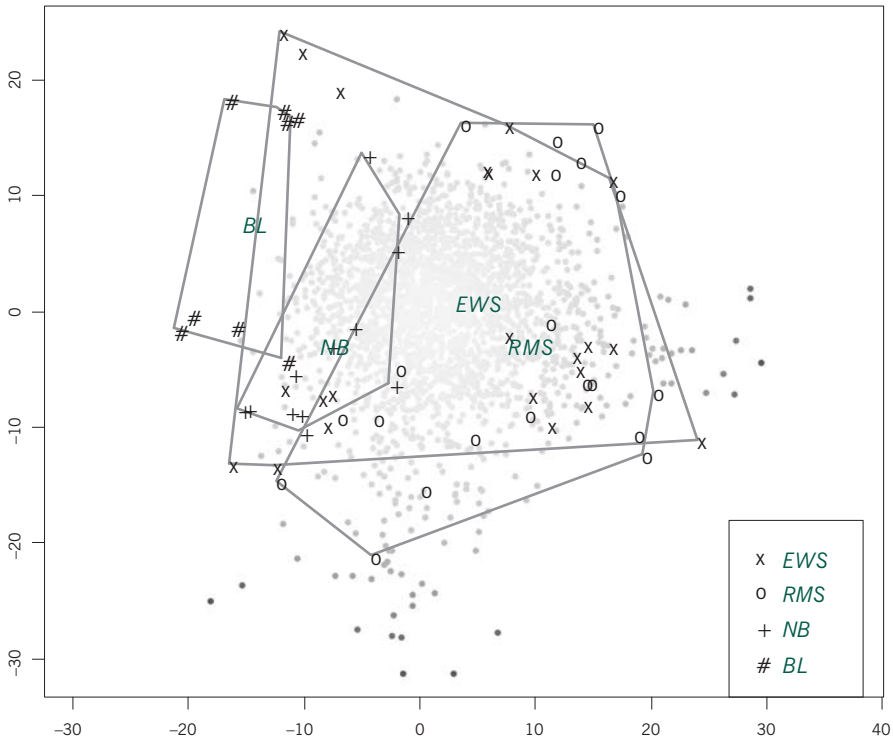
especially the relation of genes to various diseases. The samples are from small, round blue-cell tumours found in children. The genes are quantified by their expression values, the logarithm of the ratio R/G, where R is the amount of gene-specific RNA in the target sample that hybridizes to a particular (gene-specific) spot on the microarray, and G is the corresponding amount of RNA from a reference sample. The data set is called "wide" because of the large number of variables compared to the samples. The tumours fall into four major types: *EWS* (Ewing's sarcoma), *RMS* (rhabdomyosarcoma) *NB* (neuroblastoma) and *BL* (Burkitt lymphoma)—in this data set of 63 samples there are 23 *EWS*, 20 *RMS*, 12 *NB* and 8 *BL* tumours. There is an additional data set of 20 samples from these four cancer types, which we will use later in the case study to test a classification rule predicting cancer type.

Principal component biplot

The basic data are all on a logarithmic scale and do not require further standardization.

Notice that these logarithms of ratios are not log-ratios in the sense of Chapter 7, where the ratios are formed from all pairs of a set of observed variables. Because there are 2308 variables we will not use arrows to depict each one, but grey dots

Exhibit 13.1:
PCA contribution biplot of the data set "cancer", showing convex hulls around the four groups and labels at their centroids. Grey dots indicate the 2308 genes



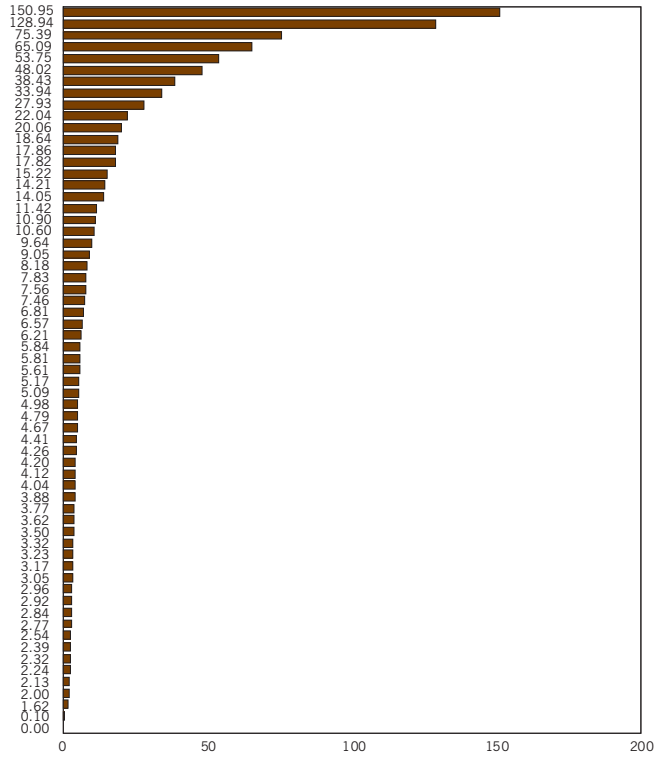


Exhibit 13.2:
Scree plot of the 63 eigenvalues in the PCA of the data set "cancer", showing the last one equal to 0 (there are 62 dimensions in this "wide" data set)

on a grey scale, where the darkness of the point is related to the gene's contribution to the solution—see Exhibit 13.1. Because this is a PCA biplot with no differential weights on the variables, the highly contributing genes will also be those far from the centre of the display.

The PCA biplot does not separate the cancer types very well, as seen by the large overlap of the four groups. Of course, this is not the objective of the PCA, which aims to maximize the between-sample dispersion, not the between-group dispersion. This sample-level biplot gives a first idea of how the samples lie with respect to one another and is useful for diagnosing unusual samples or variables, as well as spotting possible errors in the data. The dimensionality of this 63 × 2308 matrix is 62, determined by the number of samples minus 1 in this "wide" case rather than the number of variables. The percentage of variance accounted for by the two-dimensional solution is 28.5%. It is useful to look at the *scree plot* of the eigenvalues to try to assess the amount of noise in the data (Exhibit 13.2). The total variance in this data set is equal to 982.0, with an average per dimension of 982.0/62 = 15.8. By this criterion the first 14 dimensions are above average, although it is clear that the first two do separate clearly from the rest.

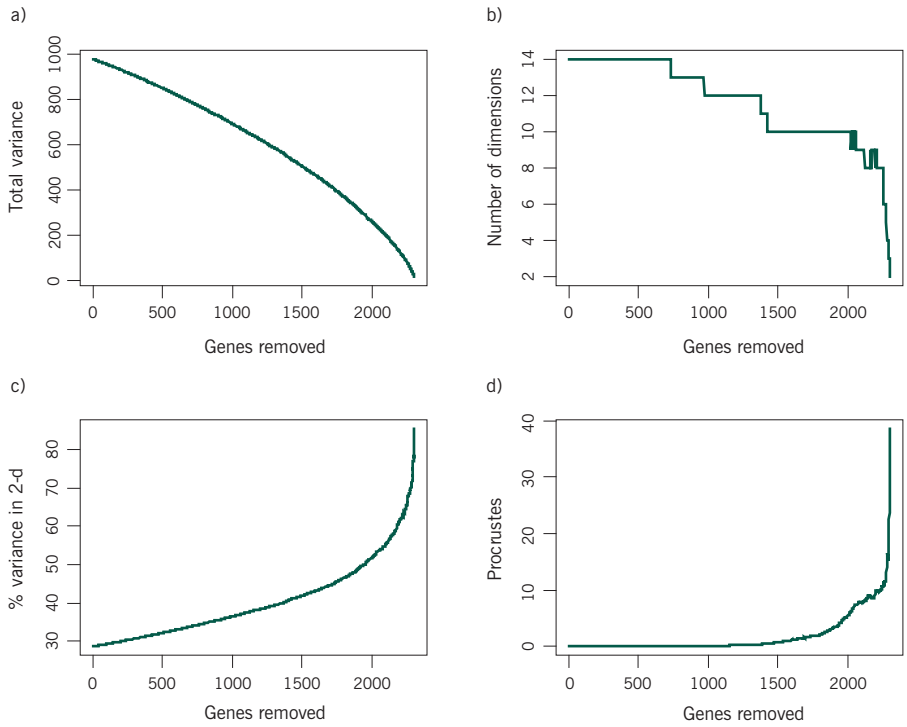
Reducing the number of variables

We have several tools at our disposition now to reduce the number of variables (genes) while keeping track of the effect this has on the visualization of the cancer samples. A possible strategy is to reduce the gene set one at a time, removing each time the gene that contributes the least to the solution. At each stage of the gene removal we measure the following aspects, shown in Exhibit 13.3:

a. The total variance and the average over the dimensions (the latter will be the former divided by 62 until the number of genes reduces below 62, in which case the dimensionality is determined by the number of genes).
b. The number of dimensions that are above the average.
c. The percentage of variance explained by the two-dimensional solution.
d. The Procrustes statistic on the configuration of sample points, compared to the initial solution (Exhibit 13.1)—this will quantify how much the configuration is changing.

Total variance (Exhibit 13.3a) obviously decreases as genes are removed—the decrease is less at the start of the process when the genes of very minor contribution to the solution are removed. The number of dimensions greater than the average also decreases (Exhibit 13.3b) but still remains fairly high until the end of the re-

Exhibit 13.3:
Monitoring of four statistics as the number of removed genes increases



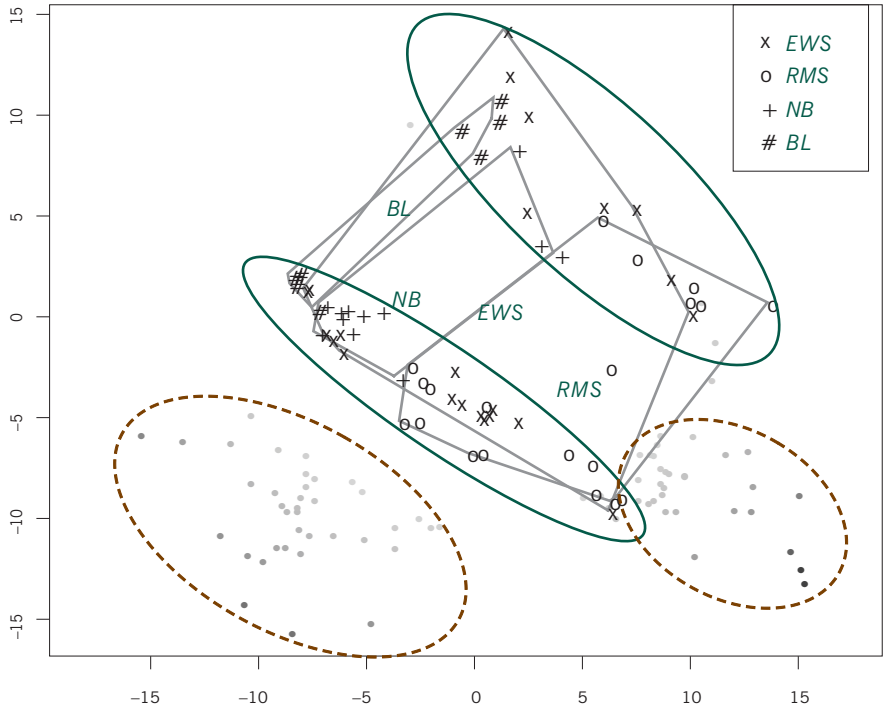


Exhibit 13.4:
PCA biplot of the reduced gene set (75 high-contributing genes, that is 2233 genes omitted), showing one set of genes (in dashed ellipse) at bottom right separating the group centroids (indicated by the labels) and another group at bottom left that is separating the total sample into two distinct groups (shown in the green ellipses), independent of their cancer types

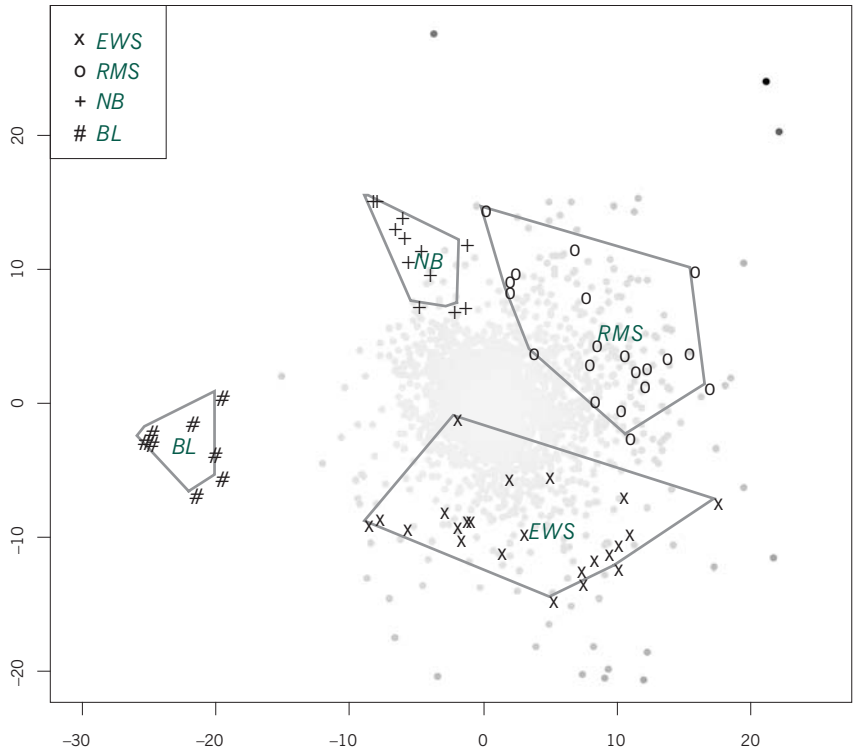
moval process. The percentage of variance on the first two axes increases as the "noisy" part of the data is removed (Exhibit 13.3c). According to the Procrustes analysis (Exhibit 13.3d) the two-dimensional configuration remains almost the same even when as many as 1500 genes are removed.

We chose a solution when the Procrustes statistic reached 10%, when 2233 genes were removed, leaving only 75 included in the PCA. Notice the gradual change in the Procrustes statistic (Exhibit 13.3d) up to this point, then a relative stability in the configuration at about 10% followed by more dramatic changes. Exhibit 13.4 shows the biplot with the reduced set of genes. The spread of the four groups, from *BL* to *RMS*, is retained (see Exhibit 13.1), just slightly rotated. What is evident here is the emergence of two groups of genes, one at bottom right which is responsible for the separation of the tumour groups, and another group at bottom left which separates the samples into two clear clusters independent of their groups—the only exception is an *RMS* tumour suspended between the two clusters.

Centroid biplot —all variables

In order to see the separation of the tumour groups better and to identify which genes are determining the difference between them, a biplot of the group centroids can be performed, as described in Chapter 11 on discriminant

Exhibit 13.5:
Centroid biplot of the four tumour groups, using all 2308 variables. The percentage of centroid variance displayed is 75.6%, with between-group variance in the plane 88.6% of the total



analysis (DA) biplots. Because there are four centroids, the space they occupy is three-dimensional; hence the planar display involves the loss of only one dimension. Exhibit 13.5 shows the centroid (or DA) biplot based on all 2308 genes.

The tumour groups are now very well separated, and the separation of the clusters observed in Exhibit 13.4 is no longer present. Of the total variance of the centroids in their full three-dimensional space, 75.6% is represented in the biplot. Of the total variance of the 63 samples represented in this two-dimensional biplot, 88.6% is between-group variance and 11.4% within-group variance.

Centroid biplot—reduced set of variables

Again, we are interested in reducing the number of genes to see which are the most determinant in separating the groups. By applying the same step-by-step reduction in the number of genes, always removing the gene with the least contribution to the group differentiation at each step, and by monitoring the percentage of variance displayed in the two-dimensional map as well as the proportion of total planar variance accounted for by the between-group part. It turns out that a

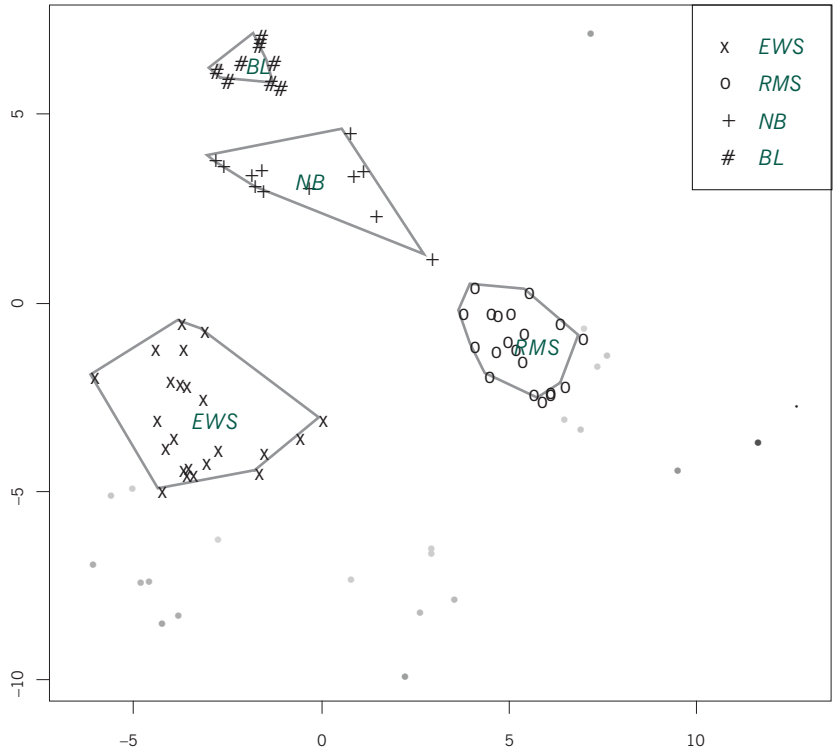


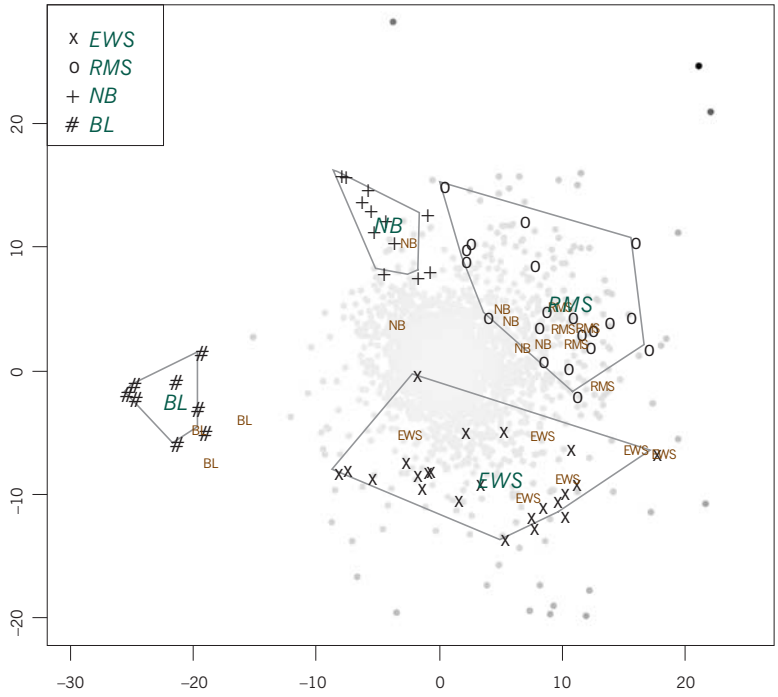
Exhibit 13.6:
Centroid biplot of the four tumour groups, using 24 highest contributing variables after stepwise removal. The percentage of centroid variance displayed is 94.9%, with between-group variance in the plane 90.5% of the total

maximum of the latter percentage is reached when we have reduced the gene set to 24 genes, for which the solution is shown in Exhibit 13.6. The between-group variance in the plane is 90.5% of the total, about 3 percentage points better than Exhibit 13.5. There is only 5.1% of the centroid variance in the third dimension now, as opposed to 24.4% in Exhibit 13.5.

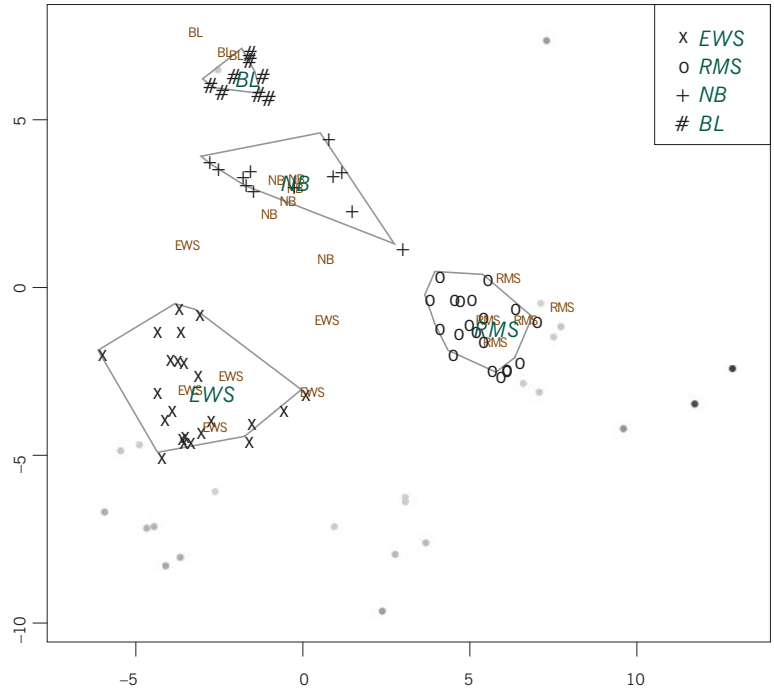
In Exhibit 13.6 we have thus achieved an optimal separation of the groups, while also reducing the residual variance in the centroids that is in the third dimension. Notice the lining up of the three tumour groups *BL*, *NB* and *RMS* from top left to bottom right, coinciding with the genes extending to the bottom right hand side: it will be these genes that distinguish these three groups, with increasing values from *BL* to *NB* to *RMS*. On the other hand the group *EWS* is situated at bottom left associated with high values of the group of genes at bottom left, and low values of the single gene that one finds at top right. There is a group of six genes at the bottom of the display that are separated from the group at bottom left, which no doubt not only separate *EWS* from the other groups but also contribute slightly to the left-to-right separation of the other three groups.

Exhibit 13.7:
The 20 additional tumours in the centroid solution space for all 2308 genes (upper biplot), and the reduced set of 24 genes (lower biplot)

a)



b)



Classification of additional samples

In addition to the 63 samples studied up to now, an additional sample of 20 tumours was available, and the type of tumour was known in each case. We can use our results in Exhibits 13.5 and 13.6 above to see whether accurate predictions of the tumour types are achieved in this test data set. We do this in a very simple way, just by situating the tumours in the two-dimensional solution space and computing their distances to the group centroids, and then predicting the tumour type by the closest centroid. Exhibit 13.7 shows the new tumours in the solution of Exhibit 13.5 using all 2308 genes (upper biplot) and then in the solution of Exhibit 13.6 using the reduced set of 24 genes (lower biplot). It is clear that in the upper biplot that four of the six *NB* tumours will be misclassified as *RMS*. In the lower biplot, the new tumours generally lie closer to their corresponding centroids, with just two *EWS* tumours being misclassified as *NB*, the one on the right being only a tiny bit closer (in the third significant digit) to the *NB* centroid than to the *EWS* one. It is a general principle that the elimination of irrelevant variables can improve the predictive value of the solution, and this is well illustrated here.

Improving prediction

As a final remark, it is possible to improve the predictive quality of this centroid classification procedure in two ways. First, there is some additional variance in the centroids in the third dimension, which we have ignored here. Calculating tumour-to-centroid distances in the full three-dimensional space of the four centroids will improve the classification. Second, in the area known as *statistical learning*, a branch of machine learning, the small subset of genes used to define the predictor space would be chosen in a more sophisticated way, using *cross-validation*. This involves dividing the *training set* of data (that is, our initial sample of 63 tumours) into 10 random groups, say, and then using 9 out of the 10 groups to determine the subset of variables that best predicts the omitted group, and then repeating this process omitting each of the other groups one at a time. There would thus be 10 ways of predicting new observations, which we would apply in turn to the *test set* (the 20 additional tumours), obtaining 10 predictions for each new tumour from which the final prediction is made by majority "vote". If these two additional improvements are implemented in our procedure it turns out that we can predict the group membership of all 20 tumours exactly.

SUMMARY

We have shown how biplots based on principal component analysis of both individual-level and aggregate-level data can be used to identify natural groups of observations in a large data set as well as distinguish between existing known groups. With respect to this data set which has a huge number of variables compared to observations:

1. In both the individual- and aggregate-level analyses, it is useful to reduce the number of variables to a smaller set that is the most determinant in showing

respectively (i) the patterns in the individual-level data, and (ii) the separation of the known groups.

2. One way of eliminating variables is to calculate each variable's contribution to the solution (a planar biplot in our application). The variable with the least contribution is eliminated, and the procedure is repeated over and over again until a small subset is found.

3. We decided to stop the variable elimination process in the individual-level analysis when the Procrustes statistic rose to 10%—this was an *ad hoc* decision, but was based on observing the evolution of the Procrustes statistic as variables were eliminated. This statistic increased very slightly and slowly up to this point, but reducing the variables beyond this stage the solution started to change dramatically

4. In the case of the aggregate-level analysis, we monitored the ratio of between-group variance to total variance in the low-dimensional solution as variables were eliminated, and stopped when this reached a maximum.

5. In the centroid analysis, the eventual space based on the smaller set of variables can be used to classify new observations, by calculating their distances in the solution to the centroids and then choosing the centroid that is closest as the group prediction.