Selection bias in plots of microarray or other data that have been sampled from a high-dimensional space

J. H. Maindonald* C. J. Burden†

(Received 16 November 2004, revised 14 February 2005)

Abstract

For data that have many more features than observations, finding a low-dimensional representation that accurately reflects known prior groupings is non-trivial. Microarray gene expression data, used to create a "signature" or discrimination rule that distinguishes cancer tissues that are classified according to type of cancer, is an important special case. The optimal number of features is suitably determined using cross-validation, in which each of several parts of the data becomes in turn the test set, with the remaining data used for training.

*Centre for Bioinformation Science, Math. Sci. Inst., Australian National University, Canberra, ACT 0200, Australia. mailto:john.maindonald@anu.edu.au

†Centre for Bioinformation Science, John Curtin School of Medical Research & Mathematical Sciences Institute, Australian National University, Canberra, ACT 0200, Australia

See http://anziamj.austms.org.au/V46/CTAC2004/Main for this article, © Austral. Mathematical Soc. 2005. Published 15 March 2005, amended March 18, 2005. ISSN 1446-8735

At each such division or "fold" of the data into a training and test set, both the selection of features and the derivation of the discriminant rule must be repeated. Use of the complete data for prior selection of features can lead to a grossly optimistic assessment of predictive accuracy and, in scatter-plot graphs that show discriminant function scores, to a spurious or exaggerated separation between groups. At each division or fold, a second versus first discriminant axis plot of test scores can be drawn. This paper presents a method for bringing these different plots, which have different choices of features and relate to different coordinate systems, into a single plot in which the configuration of points fairly reflects the accuracy of the discriminant procedure. The methodology is applicable, in principle, to use of any discriminant analysis methodology, or of ordination or multidimensional scaling, for obtaining a low dimensional graphical representation of data.

Contents

1 Introduction

Data sets from microarray experiments typically have values of each of a large number of features (expression indices, or 'genes'), for each of a small number of biological samples (observations). More than 10,000 genes are common, while the number of samples may run from one to several hundred. The present discussion assumes that there are enough samples to allow a useful classification into groups, which in our examples will be different tissue types, and discriminant rules developed. A low-dimensional view that fairly reflects the performance of a discriminant rule can draw attention to samples that seem to be misclassified, or to apparent groupings other than those used in determining the rule, or to aberrant observations.

The acute lymphoblastic leukemia (ALL) data used for illustrating the methodology has $n = 24$ samples, grouped into $g = 4$ tissue types or sources — B-cell females (6 samples), B-cell males (11 samples), T-cell females (1 sample) and T-cell males (6 samples). The data used in most of the subsequent discussion is a matrix of expression values that has dimension $m = 4190$ features by $n = 24$ observations (samples).

With 24 samples and four groups, a maximum of 20 features can be used for the discriminant analysis, with some smaller number than 20 likely to be optimum. There is a cogent case for giving preference to features that individually show the greatest discriminatory power, on the grounds that they hold information that should be retained and represented in any plot. Here the F-statistic (with 3 and 20 degrees of freedom) will be used, though noting that other statistics might be used to similar effect. The n features are chosen whose F-statistics are largest, with n chosen to give maximum predictive accuracy.

Canonical discriminant analysis, as used in this paper, is a generalization of linear discriminant analysis. It can in principle, given sufficient data, determine up to $g - 1$ axes of discrimination between g groups, yielding a

matrix of discriminant function scores $\mathbf{C}$ that has at most $g-1$ columns. The ordering of columns reflects order of effectiveness in separating groups, and most of the relevant information is often in the first few columns. The first discriminant function gives the best discrimination in a single dimension, the second function the next best discrimination subject to being uncorrelated with previous linear discriminants, and so on.

The implementation used here is set in a Bayesian framework [7, pp. 92–105], and yields posterior probabilities of membership of the several groups. Let π_j be the prior probability that an observation belongs to the jth group, by default set equal to the proportion of observations in the jth group. The Bayes rule chooses the group j for which the distance, less $-2\log(\pi_j)$, in the space of the linear discriminant functions, is a minimum. When calculations have been completed for all folds, there is a predicted group assignment for each observation to be compared with the correct assignment. The accuracy is obtained by dividing the number of correct assignments by the total number of assignments.

Use of the magnitude of the F-statistic to determine the features that will be used as discriminators involves a trade-off between the risk of omitting genuine discriminators and the risk of including features whose F-statistics are near the extreme of the empirical distribution for the null. Commonly, some features will be chosen whose F-statistics are near the extreme of the empirical distribution for the null, leading to inevitable biases for graphical representation of the data used to derive the F-statistics. The comparison between a graph derived from random normal data in the left panel of Figure 1, and from the previously mentioned ALL (acute lymphoblastic leukemia) microarray data in the right panel, illustrates the extent of the problem. The separation between groups in the left panel is clearly spurious, while that in the right panel may in part be real.

These issues are important for the use of any supervised learning approach with data where the number of features greatly exceeds the number of observations. Our methodology can be extended for use with other ap-

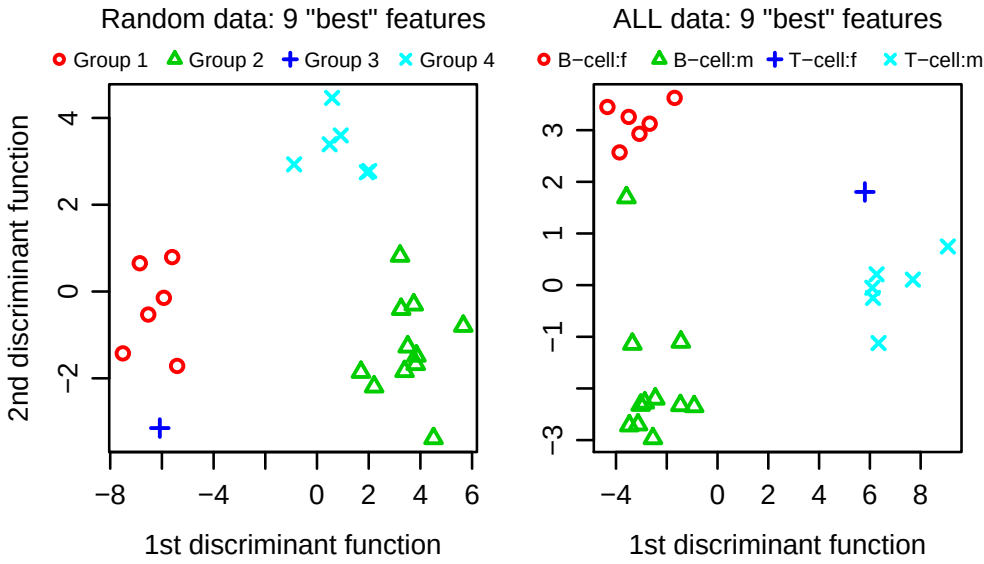

FIGURE 1: For the left panel, random normal data were placed in an array of dimension 4190 columns by 24 rows, where the rows are observations (tissue samples) and the columns are features (genes). The 24 columns were assigned to four different groups (notional "tumor types"), with frequencies 6, 11, 1 and 6 as for the microarray data that are described in the text. An anova F-test identifies the features that best separates the data into the four groups. For the right panel, the same procedure is applied to the microarray data described in the text.

proaches, though with varying complication in the ease of adaptation to give an acceptably accurate low-dimensional representation.

In [3], a subset of samples were from ALL tissues, as for the data that we have used. In that paper, the classification of ALL samples into B-type and T-type samples was based on the use of numerically based class discovery procedure that, up to that point, had recognized only the broader distinction between AML (Acute Myoblastic Leukemia) and ALL (Acute Lymphoblastic Leukemia). A two-dimensional graphical representation, such as we propose, might have been used instead. Additionally, such a graph has the potential to reveal previously unsuspected groupings, or to draw attention to misclassified samples.

Pre-processing and selection of data The 24 samples analyzed here are a subset, described as having "normal" cell genetics, from the much larger data set, described in [2], that relate to [1]. Additionally, the initial 12625 features were "filtered", without regard to the grouping into tissue types, removing features where the variation fell below a threshold. Following this filtering, 4190 features remained.

It is widely assumed that such filtering does not introduce a bias, for example, exaggerating separation between groups that are genuinely present, for use of the filtered data. Our methodology readily adapts, as will be described in the final section, to allow a check of this assumption.

2 Training/test sets, and cross-validation

Where observations (tissue samples) are split into a training set A and a test set B, the spurious clustering that is evident in the left panel of Figure 1 does not occur for a plot that shows the discriminant scores for the test data (B) alone. The crucial point is that the test data had no role in either the

TABLE 1: The table illustrates the division of a data set into four parts (one part per column) for purposes of running a cross-validation. At fold i, the scores on the training data (that is, for all except the ith part of the data) are stored in $\mathbf{Z}_i$, while the scores on the test data are stored in $\mathbf{Z}_{-i}$. At each fold, the scores and predicted group memberships for the current test data give an accurate indication of the performance of the discriminant rule.

Part 1	Part 2	Part 3	Part 4	
TEST ($\mathbf{Z}_{-1}$)	TRAIN ($\mathbf{Z}_1$)	TRAIN ($\mathbf{Z}_1$)	TRAIN ($\mathbf{Z}_1$)	Fold 1
TRAIN ($\mathbf{Z}_2$)	TEST ($\mathbf{Z}_{-2}$)	TRAIN ($\mathbf{Z}_2$)	TRAIN ($\mathbf{Z}_2$)	Fold 2
TRAIN ($\mathbf{Z}_3$)	TRAIN ($\mathbf{Z}_3$)	TEST ($\mathbf{Z}_{-3}$)	TRAIN ($\mathbf{Z}_3$)	Fold 3
TRAIN ($\mathbf{Z}_4$)	TRAIN ($\mathbf{Z}_4$)	TRAIN ($\mathbf{Z}_4$)	TEST ($\mathbf{Z}_{-4}$)	Fold 4

selection of features, or the determination of the discriminant functions and associated scores. By making B the training data and using A as the test data, a similar plot is obtained that is now limited to the data in A. The two plots will use different features and different discriminant functions, and cannot be simply superposed. Our method combines them into a single plot, albeit in the context of the more general cross-validation approach, where data are split in k parts.

For each $i = 1, \ldots, k$ in turn, the ith part becomes the test data, with the remaining data used for training. Table 1 summarizes the steps that would be followed, though for simplicity of presentation with $k = 4$ rather than with the more usual $k = 10$. A further possibility is to use leave-one-out cross-validation, so that our data would have $k = 24$. At each fold, scores, in as many dimensions as are required, will be calculated for the test data for that fold.

Note that there are two steps in the development of a discriminant rule:

- Select the features that give the best discrimination;

- Determine a discrimination rule.

The cross-validation must take account of both steps in this process, that is, the selection of features must be repeated at each fold. As a check on the effect of any filtering or other pre-processing steps, these may also be repeated at each fold.

At the ith fold, there are two sets of scores — scores $\mathbf{Z}_i$ that if used to assess predictive accuracy or plotted will give a biased assessment of performance, and scores $\mathbf{Z}_{-i}$ that give a fair assessment. A remaining task, that will be addressed below, is to approximate the scores $\mathbf{Z}_{-i}$ in a common global coordinate system, allowing all scores to be plotted on the one graph. With n_i test observations in the ith fold, the matrix $\mathbf{Z}_i$ will be $n - n_i$ by p and $\mathbf{Z}_{-i}$ will be be n_i by p, where p may be taken to be 2 or 3.

At each fold the procedure predicts, for observations in the test data, the group to which the observation should be assigned. When calculations have been completed for all folds, there is a predicted group assignment for each observation to compare with the correct assignment. Obtain the accuracy by dividing the number of correct assignments by the total number of assignments.

The optimal number of features is determined by repeating the cross-validation procedure for each of a range of numbers of features, then choosing the number that gives the greatest predictive accuracy. In the present instance, where the number of observations is small and there is substantial variation from one cross-validation run to another, the cross-validation procedure is repeated five times for each choice of number of genes, with a different random split of the observations into $k = 10$ parts at each run.

Figure 2 (blue points and curve) summarizes results. Failure to reselect features at each cross-validation fold gives the clear bias shown by the gray points and fitted curve. The resubstitution measure (in red), obtained by using all data set both for training and testing, must inevitably increase

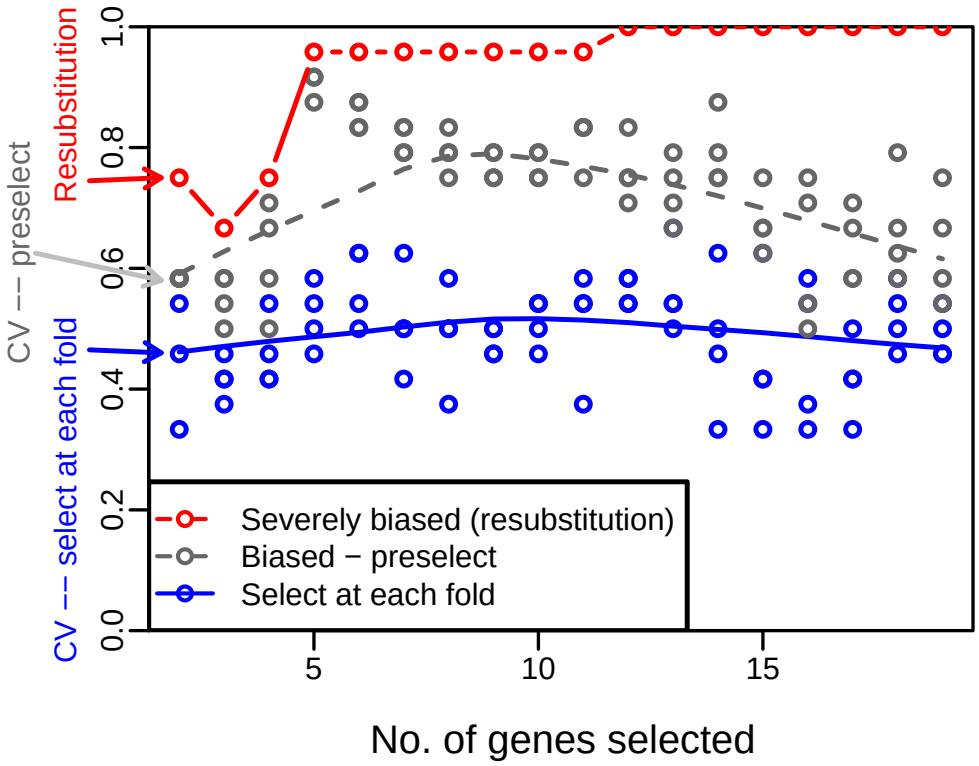

FIGURE 2: The blue points and curve are from the full cross-validation (CV) accuracy assessments. The fitted spline curve suggests that the optimum choice of number of features is 8 or 9, though accuracy does not vary greatly in a range of 2 or 3 either side of that figure. For the gray points and curve, and the red points, see text.

as the number of features is increased. Figure 1 corresponds to the red point for $n = 9$. Thus if Figure 1 accurately reflected the discriminatory power, canonical discriminant analysis would give a rule that has the grossly optimistic 96% predictive accuracy that is shown for the red point at $n = 9$ in Figure 2.

3 Approximation of test scores in a common coordinate system

The idea is to use the scores on the training data, for each fold, in order to make a connection with scores that are derived for the data as a whole. There are several ways that such global or common scores might be derived.

- Average over the scores that have been derived for the k folds.
- Replace the m features by at most n sets of principal component scores, use these as data for a discriminant analysis that uses all observations, and calculate discriminant function scores.
- Using all observations, select the subset of features that have the largest F-statistics, and use these for a discriminant analysis. Scores from this analysis are then used as the global scores.

However derived, the global score matrix will be written $\mathbf{G}$. The matrix that holds scores for the subset of observations that are included in the training set for the ith fold will be written $\mathbf{G}_i$. Thus the rows of the matrices $\mathbf{G}_i$ and $\mathbf{Z}_i$ relate to the same observations.

Calculations proceed by finding, for each $i = 1, \ldots, k$, the transformation $\mathbf{M}_i$ such that $\mathbf{Z}_i\mathbf{M}_i$ best approximates $\mathbf{G}_i$ in a least squares sense, that

is, $\mathbf{M}_i$ is chosen to minimize the trace of

$$(\mathbf{G}_i - \mathbf{Z}_i\mathbf{M}_i)'(\mathbf{G}_i - \mathbf{Z}_i\mathbf{M}_i)\,.$$

Next, for each $i = 1, \ldots, k$, calculate

$$\mathbf{C}_i = \mathbf{Z}_{-i}\mathbf{M}_i\,,$$

where $\mathbf{C}_i$ is an n_i by p matrix. Merging the rows of the $\mathbf{C}_i$ $(i = 1, \ldots, k)$ gives an n by p matrix $\mathbf{C}$. Columns of $\mathbf{C}$ are then, in order, "global" cross-validation scores for the first, second, ..., discriminant functions, and pairwise plots are made as required.

Examples of the use of the methodology

Figure 3 essentially reproduces Figure 1, but here using the cross-validation scores. The left panel now shows, correctly, no discrimination between the groups. More importantly, the right panel shows, by comparison with Figure 1, very little discrimination between groups.

Figure 4 is for data where our procedure shows a clear separation into groups. The interest lies in whether the ALL tissues that were used in [3] could be further subdivided according to the sex of the patient and the source of the tissue — bone marrow or peripheral blood.

4 Commentary and extensions

Clearly, the linear transformations $\mathbf{M}_i$ introduce unwanted and perhaps unavoidable noise into the global positioning of points.

An important question is whether the filtering that determines the reduced set of features used for the analysis may itself bias predictive accuracy. This can be checked by working with the total set of 12165 features

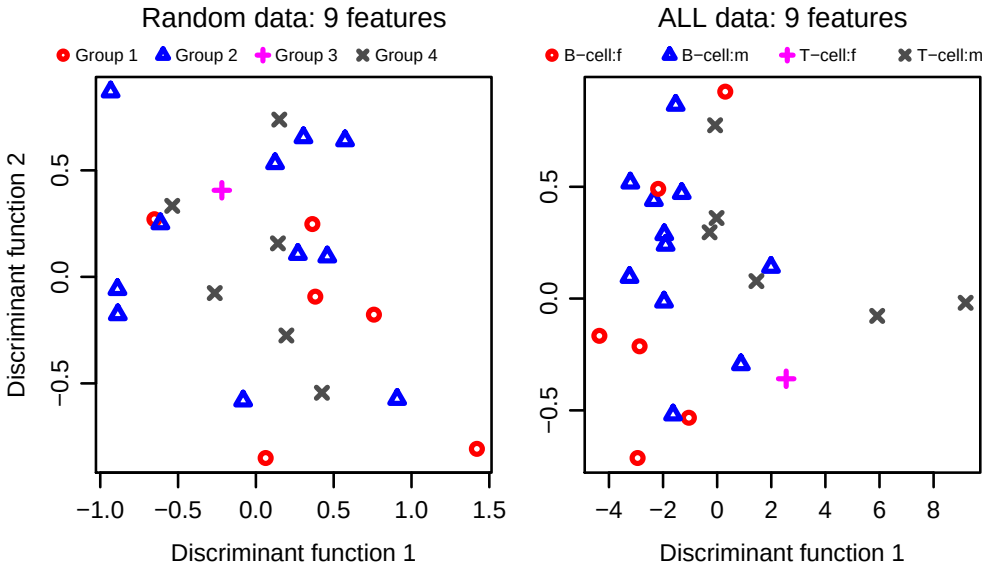

FIGURE 3: Graphs are for the same data as described in Figure 1. Our cross-validation procedure is followed, with the local discriminant scores transformed back to global discriminant axes, to give the columns of the matrix **C**. The plot on the left is for random data, whereas the plot on the right is for the same ALL microarray data as in Figure 1.

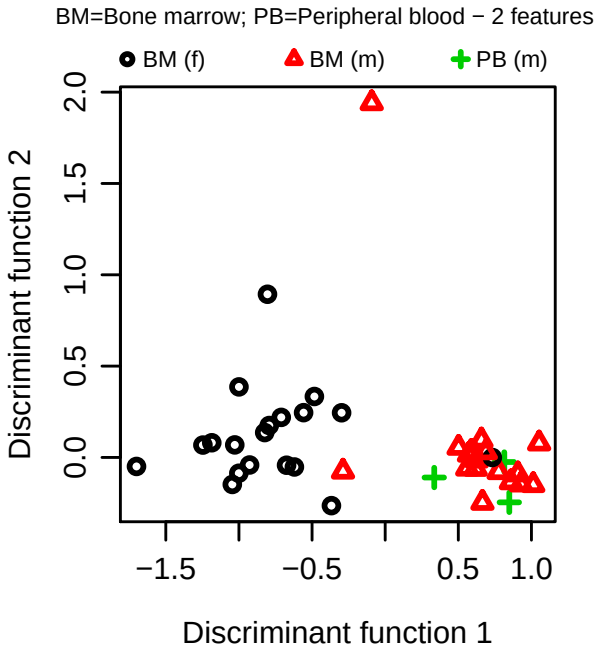

FIGURE 4: For the ALL data [3] that were the basis for this graph, there is a clear separation into two groups, with one sample that stands apart.

and repeating the filtering procedure at each fold of the cross-validation. It is sufficient to check that the filtering has no effect on the choice of features at the several folds of the cross-validation.

The matrices $\mathbf{M}_i$ represent affine transformations. An alternative is to combine a rigid transformation with shift and dilation, using Procrustes transformations [8]. With the relatively small number of observations used in this paper, an affine transformation seems to give a better approximation. Details will be given elsewhere.

The approach readily generalizes to any technique that yields a ranked set of columns of discriminant function scores, as for canonical discriminant analysis. Support Vector Machines are not obviously designed to allow a low-dimensional representation, so that their use in this context requires adaptation. Where optimal discrimination requires more than two or three features, multi-dimensional scaling or an ordination technique makes it possible to approximate results in a low-dimensional space, though with the caveat that any low-dimensional representation risks loss of information.

Principal components and other ordination techniques are likewise prone to discrimination based selection effects, though as there is no attempt to choose axes that optimally exhibit the separation between groups, the bias should be less extreme. The present methodology can be adapted for use with these methods also. With some extension, it might be used with the biplot methodology [6]. The first two or three principal components, for the total data, determine the coordinate system that will be used for the global graphical representation. Principal components results from the successive folds of the cross-validation are represented in this global coordinate system.

Acknowledgments: We thank Yvonne Pittelkow and Susan Wilson for helpful comments. John Maindonald's research was supported by ARC grant DP0343727.

A Computer implementation

We use the R system [4]. Principal components calculations use functions that are available in R for the singular value decomposition. Discriminant function calculations use the function `lda()` from Venables and Ripley's MASS package for R. For obtaining the transformations $\mathbf{M}_i$, we use the R function `qr.solve()`. An R *Sweave* [5] file that may be used to reproduce the calculations, graphs and associated commentary will be posted on the web, at http://www.maths.anu.edu.au/~johnm/r/cvplot

References

[1] Chiaretti, S., Xiaochun Li, Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz, J. and Foa, R. Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood* 103(7), 2004. C64

[2] Xiaochun Li ALL: A data package. R package version 1.0.2, 2004. [Online] http://www.bioconductor.org/data/experimental/html/ALL.html C64

[3] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield. C. D., and Lander, E. S. 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286: 531–537. C64, C69, C71

[4] Ihaka, R. and Gentleman, R. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 5: 299–314, 1996. C73

[5] Leisch F. Sweave User Manual. [Online] http://www.ci.tuwien.ac.at/~leisch/Sweave. C73

[6] Pittelkow, Y. E., Wilson, S. R. Visualisation of Gene Expression Data: The GE-biplot, the Chip-plot and the Gene-plot. *Statistical Applications in Genetics and Molecular Biology* (19pp). [Online] http://www.bepress.com/sagmb/vol2/iss1/art6, 2003. C72

[7] Ripley, B. D. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996. C62

[8] Sibson, R. Studies in the robustness of multidimensional scaling: Procrustes analysis. *Journal of the Royal Statistical Society B* 40: 234–238, 1978. C72