Cluster validation by prediction strength

Robert Tibshirani, *Guenther Walther †
David Botstein,‡and Patrick Brown§

September 1, 2001

Abstract

We propose a new quantity for assessing the number of groups or clusters in a dataset. The key idea is to view clustering as a supervised classification problem, in which we must also estimate the "true" class labels. The resulting "prediction strength" measure assesses how many groups can be predicted from the data, and how well. In the process, we develop novel notions of bias and variance for unlabelled data. Prediction strength performs well in simulation studies, and we apply it to clusters of breast cancer samples from a DNA microarray study. Finally, some consistency properties of the method are established.

1 Introduction

Cluster analysis is an important tool for "unsupervised" learning— the problem of finding structure in data without the help of a response variable. A major challenge in cluster analysis is estimation of the appropriate number of groups or clusters. Many existing methods for this problems focus on the

*Division of Biostatistics and Department of Statistics, Stanford University, Stanford CA 94305; tibs@stat.stanford.edu

†Department of Statistics, Stanford University, Stanford CA 94305; walther@stat.stanford.edu

‡Department of Genetics, Stanford University;botstein@genome.stanford.edu

§Department of Biochemistry, Stanford University;pbrown@cmgm.stanford.edu

within-cluster dispersion W_k, resulting from a clustering of the data into k groups. The error measure W_k tends to decrease monotonically as the number of clusters k increases, but from some k on the decrease flattens markedly. Statistical folklore has it that the location of such an "elbow" indicates the appropriate number of clusters.

A number of methods have been proposed for estimating the number of clusters, some of which exploit this elbow phenomenon. Many proposals are summarized in the comprehensive survey in Milligan & Cooper (1985), while Gordon (1999) discusses the best performers. More recent proposals include Tibshirani et al. (2001), Sugar (1998) and Sugar et al. (1999). However it is not clear if these methods are widely used: this may be because they are difficult to intrepret.

In this paper we take a different approach. We view estimation of the number of clusters as a model selection problem. Now in classification problems with labelled data, model selection is usually done by minimization of prediction error. This is compelling, and also provides an estimate of the prediction error for individual observations. Here we develop a corresponding method for estimating the number of clusters by adapting prediction ideas to clustering. By focussing on prediction error rather than the within sum of squares W_k, the results of the procedure are directly interpretable and information about the cluster membership "predictability" of individual observations is available.

Section 2 describes the basic procedure for estimating prediction strength. In Section 3 we give background motivation for the method. We define new notions of bias, variance and prediction error for clustering, and show that prediction strength essentially estimates the variance term. Section 4 examines how well the procedure estimates the "true" prediction strength. Up to this point, K-means clustering is the focus. In Section 5 we discuss application of the technique to hierarchical clustering. Section 6 describes a simulation study, comparing the method to other competing methods for estimating the number of clusters. Finally in Section 7 we establish consistency of the method in a simple but informative case.

2 Prediction strength of clustering

Our training data $X_{tr} = \{x_{ij}\}, i = 1, 2, \ldots n;, j = 1, 2, \ldots p$ consist of p features measured on n independent observations. Let $d_{ii'}$ denote the distance between observations i and i'. The most common choice for $d_{ii'}$ is the squared Euclidean distance $\sum_j (x_{ij} - x_{i'j})^2$.

Suppose we cluster the data into k clusters. For example we might use k-means clustering based on Euclidean distance, or hierarchical clustering. Denote this clustering operation by $C(X_{tr}, k)$.

Now when we apply this clustering operation to the training data, each pair of observations either does or does not fall into the same cluster. To summarize this, let $D[C(\ldots), X_{tr}]$ be an $n \times n$ matrix, with ii'th element $D[C(\ldots), X_{tr}]_{ii'} = 1$ if observations i and i' fall into the same cluster, and zero otherwise. We call these entries "co-memberships".

Our proposal for real data uses repeated cross-validation. To motivate this approach, consider the conceptually simpler scenario in which an independent test sample X_{te} of size m is available, drawn from the same population as the training set. As above, we can cluster X_{te} into k clusters via an operation $C(X_{te}, k)$, and summarize the cluster co-memberships via the $m \times m$ matrix $D[C(X_{te}, k), X_{te}]$.

The main idea of this paper is to a) cluster the test data into k clusters; b) cluster the training data into k clusters, and then c) measure how well the training set cluster centers predict co-memberships in the test set. For each pair of test observations that are assigned to the same test cluster, we determine whether they are also assigned to the same cluster based on the training centers.

Figure 1 illustrates the this idea. The data lie in two clusters. In the top row 2-means clustering is applied to both the training and test data. In the top right panel, the training centroids classify the test points into the same two green and red clusters that appear in the middle panel. But in the bottom row when three centroids are used, the classifications by test and training centroids differ substantially. Here is the idea in detail. For a candidate number of clusters k let $A_{k1}, A_{k2}, \ldots A_{kk}$ be the indices of the test observations in test clusters $1, 2, \ldots k$. Let $n_{k1}, n_{k2}, \ldots n_{kk}$ be the number of observations in these clusters.

We define the "prediction strength" of the clustering $C(\cdot, k)$ by

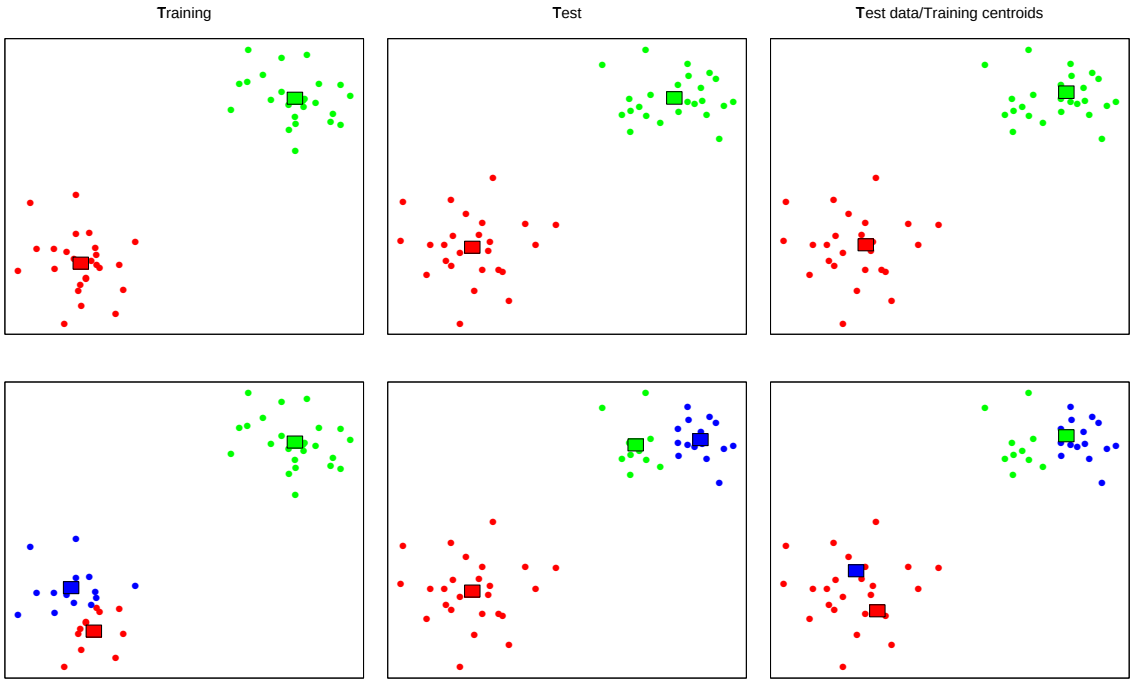

Figure 1: *Illustration of prediction strength idea. Data is simulated in two well-separated clusters. In the top row k-means clustering with two centroids is applied to both the training and test data. In the top right panel, the training centroids classify the test points into the same two green and red clusters that appear in the middle panel. However in the bottom row, when three centroids are used, the classifications by test and training centroids differ considerably.*

$$\text{ps}(k) = \min_{1 \le j \le k} \frac{1}{n_{kj}(n_{kj}-1)} \sum_{i \ne i' \in A_{kj}} I(D[C(X_{tr}, k), X_{te}]_{ii'} = 1). \qquad (1)$$

For each test cluster, we compute the proportion of observation pairs in that cluster that are also assigned to the same cluster by the training set centroids. The prediction strength is the minimum of this quantity over the k test clusters.

Here is the intuition behind this idea. If $k = k_0$, the true number of clusters, then the k training set clusters will be similar to the k test set clusters, and hence will predict them well. Thus $\text{ps}(k)$ will be high. Note that $\text{ps}(1) = 1$ in general, since both the training and test set observations all fall into one cluster. However when $k > k_0$, the extra training set and test set clusters will in general be different, and thus we expect $\text{ps}(k)$ to be much smaller. We could use an average rather than minimum in expression (1); we have found that use of the minimum makes the procedure more sensitive in many-cluster situations, in accordance with the theory developed in Section 7.

Note that in general it would be difficult to compare the training and test clusterings by associating each of the k training clusters with one of the test clusters. By focussing only on the pairwise co-memberships in (1), we finesse this problem. The identity of the cluster containing each observation is not considered: only its co-memberships in *some* cluster are used.

Figure 2 shows examples with one, two and three clusters. This and other experiments suggest that we choose the optimal number of clusters $\hat{k}$ to be the largest k such that $\text{ps}(k)$ is above some threshold. Experiments reported later in the paper show that a threshold in the range $0.8 - 0.9$ works for well separated clusters. We think of $\hat{k}$ as the largest number of clusters that can be accurately predicted in the dataset.

Now in the absence of a test sample, we instead use repeated r-fold cross-validation to estimate the prediction strength (1). The first $r - 1$ folds represent the training sample, while the last fold is the test sample. In experiments reported in Section 4 we investigate 2-fold and 5-fold cross-validation. Their performance is quite similar, and we settle on 2-fold cross-validation for the rest of the paper.

Prediction strengths for individual observations can also be defined. Specif-

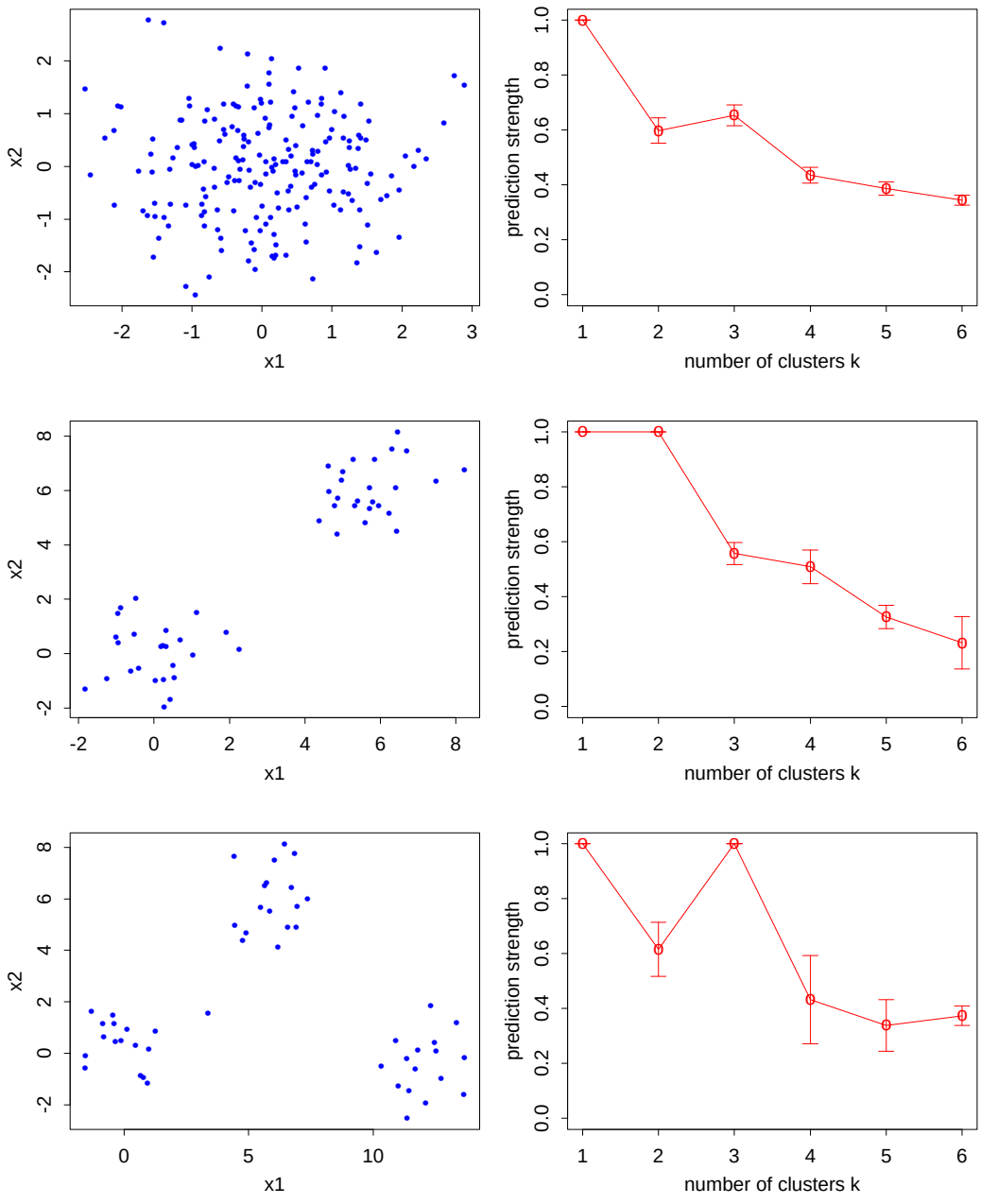

Figure 2: *Results for one, two and tree cluster examples. The test data are on the left, and prediction strength on the right. The vertical bars on the right give the standard error of the prediction strength over 5 cross-validation folds.*

ically, we define the prediction strength for observation i as

$$\mathrm{ps}(i,k) = \frac{1}{\#A_k(i)} \cdot \sum_{i' \in A_k(i)} I(D[C(X_{tr},k), X_{te}]_{ii'} = 1) \tag{2}$$

where $A_k(i)$ are the observations indices i' such that $i \neq i'$ and $D[C(X_{te},k), X_{te}]_{ii'} = 1$. Figure 3 shows an example with two centroids fit to two fairly well-separated clusters. The red points have prediction strength greater than .90, while the green points lying in the overlap region have lower prediction strength (marked on the plot).

3 Bias, variance and prediction strength for clustering

In this section we provide background motivation for the prediction strength idea. In the process we formulate novel notions of bias, variance and prediction error for clustering, analogous to the definitions for supervised learning.

Let $C^*(X)$ denote the true grouping of the data X, i.e. $D[C^*(X), X]_{ij} = 1$ iff $\underline{X}_i$ and $\underline{X}_j$ are from the same group. Define the prediction error (loss) of the clustering procedure C by

$$err_C(k) = \frac{1}{n^2} \sum_{i,j=1}^{n} \left| D[C^*(X), X]_{ij} - D[C(X,k), X]_{ij} \right|. \tag{3}$$

As all matrix entries are either 0 or 1, one sees that $err_C(k)$ decomposes into two parts:

$$err_C(k) = \begin{pmatrix} \text{proportion} \\ \text{of pairs } (\underline{X}_i, \underline{X}_j) \text{ that} \\ C(X,k) \text{ erroneously as-} \\ \text{signs to the same group} \end{pmatrix} + \begin{pmatrix} \text{proportion} \\ \text{of pairs } (\underline{X}_i, \underline{X}_j) \text{ that} \\ C(X,k) \text{ erroneously as-} \\ \text{signs to different groups} \end{pmatrix} \tag{4}$$

The first term tends to decrease as k increases (as fewer groups are erroneously aggregated into one big group), while the second term tends to increase with k (as more groups are erroneously split up into several groups). Thus the two terms have the analogous qualitative behavior of the bias and variance terms of a prediction error when the smoothing parameter is varied.

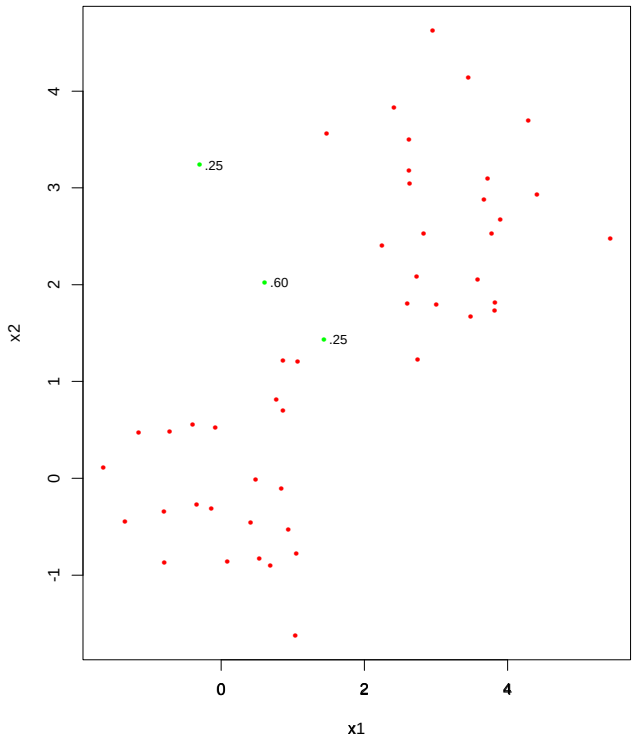

Figure 3: *Individual prediction strengths, when the data shown is clustered into 2 clusters. Green:* ps $< .90$ *(prediction strength indicated); Red:* ps $> .9$. *We used the test sample shown, and 5 randomly generated training samples from the same population. The predictions strengths were estimated from averages over the 5 training samples.*

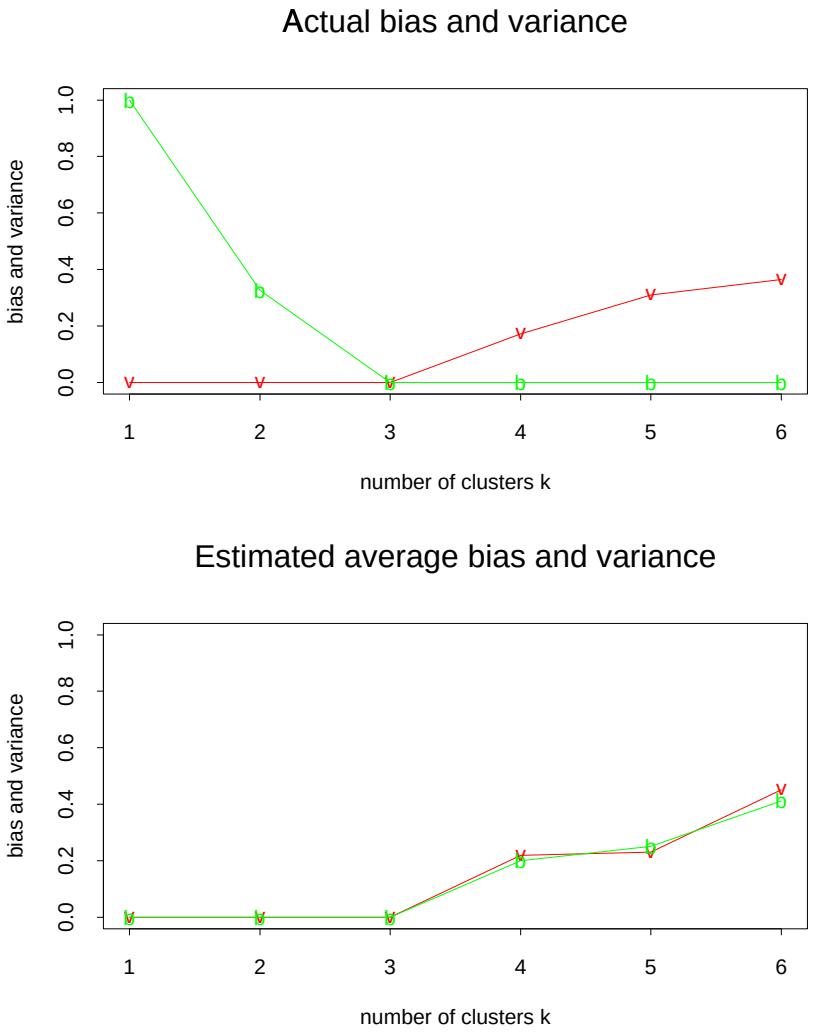

Figure 4: *Bias and variance for the 3 cluster example in Figure 2. The top panel shows the actual bias and variance using the real class labels, as in (4). The bottom panel uses $C(X_{te}, k)$ in place of the true labels $C^*(X)$.*

We can try to mimic this decomposition to estimate k, by letting $C(X_{te}, k)$ and $C(X_{tr}, k)$ take the roles of $C^*(X)$ and $C(X, k)$, respectively.

The resulting estimate of variance in the bottom panel of Figure 4 is a reasonable approximation to the variance in the top panel, but this is not the case for the estimate of bias. Substitution of $C(X_{te}, k)$ in place of the true labels $C^*(X)$ leads to a poor estimate of bias when k is less than the true number of clusters. Although we would ideally like to estimate prediction error, we instead restrict attention to the variance— the only component we can estimate well. Rather than seek the minimum point of prediction error, we seek the point at which the variance starts to rise significantly. Prediction strength, defined above, equals one minus the variance.

If the trial value k is chosen too large, then we expect the variance in at least one cluster to be significantly larger than zero. Thus we consider the worst performance of the procedure among the k clusters and hence seek k to minimize

$$\max_{1\le j\le k} \frac{1}{n_{kj}(n_{kj}-1)} \sum_{i\neq i'\in A_{kj}} 1\Big(D[C(X_{tr},k), X_{te}]_{ii'} = 0\Big) \tag{5}$$

or alternatively, we choose k to maximize the *prediction strength*

$$\text{ps}(k) = \text{cv-ave} \min_{1\le j\le k} \frac{1}{n_{kj}(n_{kj}-1)} \sum_{i\neq i'\in A_{kj}} 1\Big(D[C(X_{tr},k), X_{te}]_{ii'} = 1\Big)$$

where we modified the preliminary definition (1) by averaging over several random splits of the data into X_{te} and X_{tr}, denoted by cv-ave. Thus, for each test set cluster j, we compute the proportion of pairs in A_{kj} that are assigned to the same group by the training set based clustering. We estimate the number of groups $\hat{k}$ in X by the largest k that maximizes $\text{ps}(k)$; taking $\hat{k}$ to be the largest k such that $\text{ps}(k) \geq 0.8$ or 0.9 works well in practice. We think of $\hat{k}$ as the largest number of clusters that can be accurately predicted in the dataset.

4 Effects of reduced sample size

There is a potential problem in using two-fold (or other r-fold) cross-validation in estimating prediction strength. With $n = 100$ observations say, two-fold

cross-validation uses training sets of size 50. The prediction strength for $n = 50$ is probably lower than that for $n = 100$, and hence our estimate will tend to be biased downward. Here we investigate this bias, and also consider 5-fold cross-validation as an alternative strategy.

We need some additional notation. Let ps_{n_1,n_2} be the prediction strength using training and test sets of size n_1 and n_2, respectively. Then given a training set of size n, the "true" prediction strength is $\text{ps}_{n,\infty}$ while two-fold cross-validation estimates this quantity using $\text{ps}_{n/2,n/2}$.

We carried out a simulation study to assess the error $\text{ps}_{n/2,n/2} - \text{ps}_{n,\infty}$. The data were generated in two standard Gaussian classes, with independent components in d dimensions, $d = 1, 5, 1000$. The first class has its centroid at the origin. For $d = 1, 5$ the second class is shifted by an amount Δ, with Δ taking values $3, 2, 1, .75, .5, .25$. For $d = 1000$, 5% of the data are randomly selected, and only those centroid components are shifted by Δ. With $n = 50$, the left panel of Figure 5 shows the error $\text{ps}_{n/2,n/2} - \text{ps}_{n,\infty}$ plotted as a function of the "true" prediction strength $\text{ps}_{n,\infty}$. In the right panel we assess 5-fold cross-validation, and hence we have plotted $\text{ps}_{4n/5,n/5} - \text{ps}_{n,\infty}$. In each case we show the mean over 10 simulations, with one standard-error bands. There appears to be no advantage in using 5-fold over 2-fold cross-validation, and hence we used the latter in this paper.

5 Applications to hierarchical clustering

One of the main motivations for this work is the widespread use of hierarchical clustering for DNA microarrays. Clustering of gene expression profiles is often used to try to discover subclasses of disease. Validation of these clusters is important for accurate scientific interpretation of the results.

Figure 6 shows a dendrogram from hierarchical clustering of the gene expression of 85 breast cancer patients. This data is taken from (Perou et al. 1999). In these applications the hierarchical clustering is performed "bottom-up", starting with individual samples, and agglomerating them. The dendrogram in figure 6 is plotted upside down relative to the usual plot, so the individual samples are actually at the top. The study of (Perou et al. 1999) discovered at least four interesting classes of breast cancer, labelled in the dendrogram. [1] Hierarchical clustering is preferred to K-means

[1] This example is for illustration purposes. Our dendrogram is not the same as that of

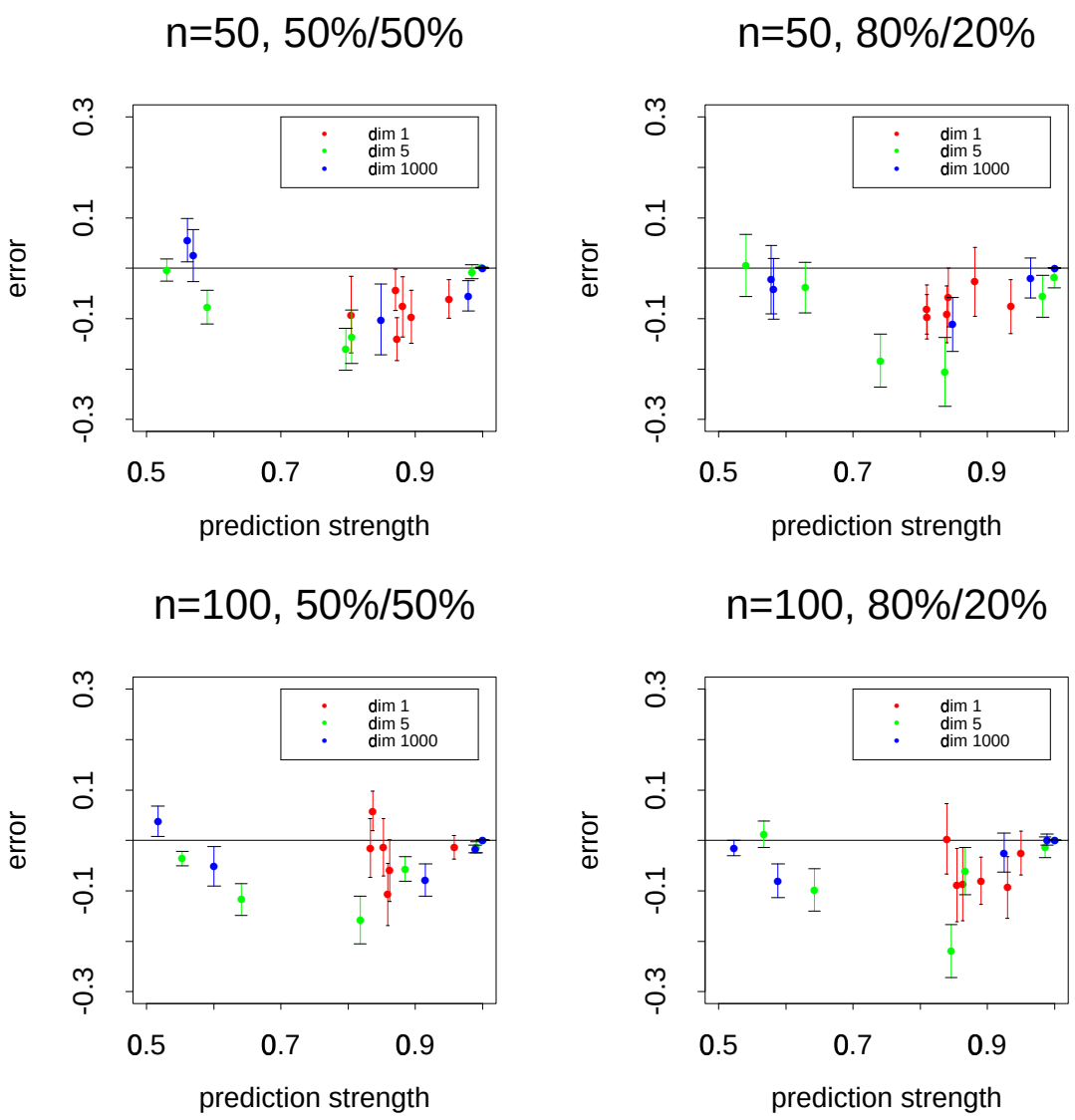

Figure 5: *Bias in prediction strength estimates, for the experiments described in the text. The left panel shows the error* $\mathrm{ps}_{n/2,n/2} - \mathrm{ps}_{n,\infty}$ *plotted as a function of the "true" prediction strength* $\mathrm{ps}_{n,\infty}$*. The right panel assesses 5-fold cross-validation, and hence* $\mathrm{ps}_{4n/5,n/5} - \mathrm{ps}_{n,\infty}$ *is shown.*

Table 1: *Prediction strength for all pairs of the 4 groups from Figure 6. The last column contains the averages for each row.*

	Lum B/C	Lum A	Normal	Basal	Average
Lum B/C		.80	.92	.71	.81
Lum A	.80		.59	.77	.72
Normal	.92	.59		.62	.71
Basal	.71	.77	.64		.71

in this context, because it shows the whole spectrum of different K all in the same picture.

The question that arises is: how different are these four groups? To help answer this, we can apply the prediction strength idea, for example to study the two main branches in Figure 6. One could apply hierarchical clustering to define the clustering operation $C(X_{tr}, k)$, cutting off the resulting dendrogram at a height that produces k clusters. However this does not seem appropriate, as hierarchical clustering is performed bottom-up, and the resulting k groups might look nothing like the original ones. Rather we adapt the following strategy: use hierarchical clustering to find potential clusters as in Figure 6, but then use the k-means clustering as $C(X_{tr}, k)$ in the calculation of prediction strength. k-means clustering is a top-down method, and is better suited to finding large groups.

Using this idea, we can estimate the prediction strength of any two-class division in the dendrogram. In Figure 6, we have labelled the splits at the first two levels with the estimated prediction strength. For example, the (Luminal B/C and A) versus (Normal and Basal/ERBB2) has a prediction strength of only .58. We can look deeper by computing the prediction strength of all pairs of the 4 groups by using only the corresponding data. The results are given in Table 1. We see that the luminal B/C group is well separated, especially from the Normal group. Most other pairs are not that well separated.

(Perou et al. 1999). These authors used a non-standard form of average linkage clustering applied to a selected list of 450 genes. We did not have easy access their algorithm, so instead used the standard clustering procedure in S-PLUS. We also used a smaller subset of genes, so that our dendrogram looked roughly like theirs.

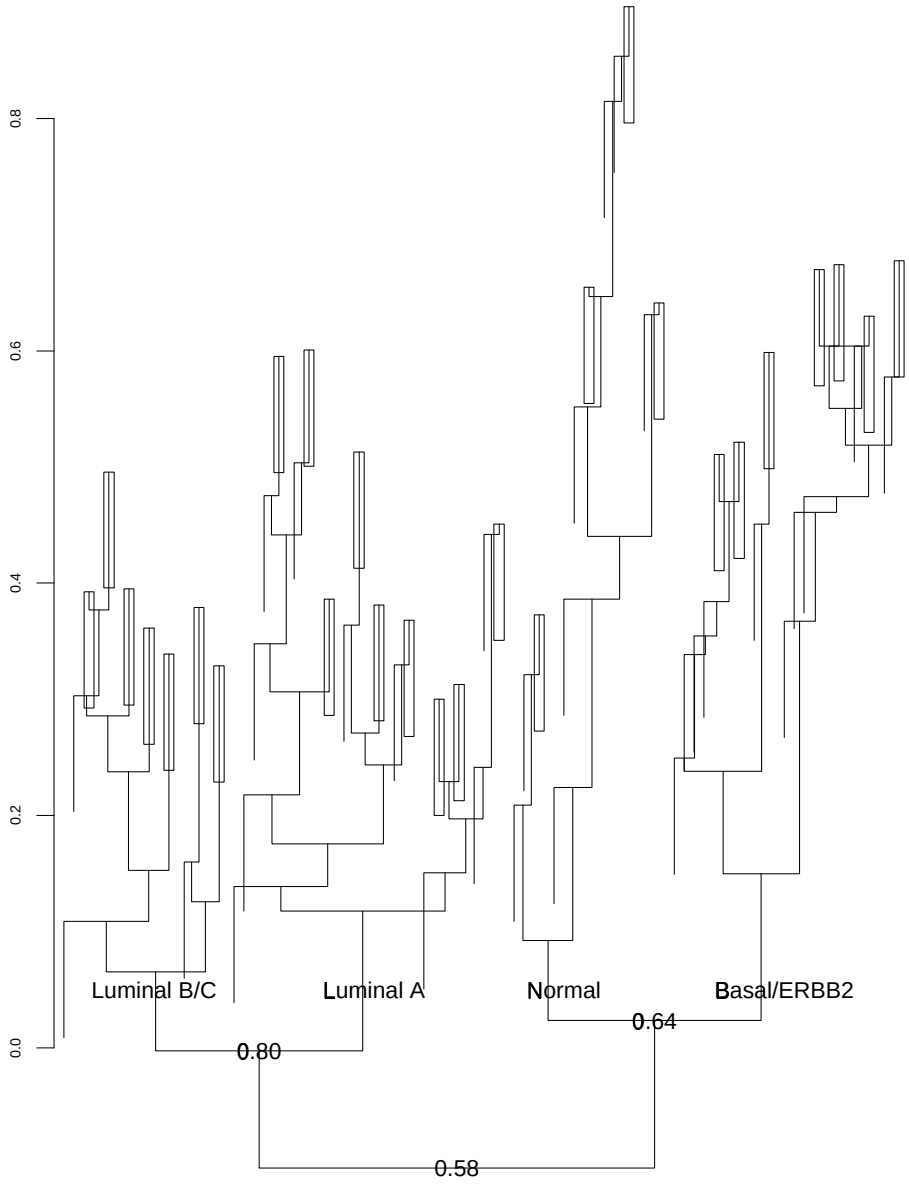

Figure 6: *Dendrogram from breast cancer study, with the estimated prediction strength at the upper branches.*

6 A simulation study

In this section we replicate the simulation study done in Tibshirani et al. (2001), comparing a number of different methods for estimating the number of clusters. We now include the prediction strength method in the comparison.

We generated datasets in five different scenarios:

1. *Null (single cluster) data in ten dimensions:* 200 data points uniformly distributed over the unit square in ten dimensions

2. *3 clusters in 2 dimensions*: the clusters are standard normal variables with (25, 25, 50) observations, centered at (0,0), (0,5), and (5,-3)

3. *4 clusters in 3 dimensions*: each cluster was randomly chosen to have 25 or 50 standard normal observations, with centers randomly chosen as $N(0, 5 \cdot I)$. Any simulation with clusters having minimum distance less than 1.0 units between them was discarded.

4. *4 clusters in 10 dimensions*: each cluster was randomly chosen to have 25 or 50 standard normal observations, with centers randomly chosen as $N(0, 1.9 \cdot I)$. Any simulation with clusters having minimum distance less than 1.0 units between them was discarded. In this and the previous scenario, the settings are such that about one-half of the random realizations were discarded.

5. *Two elongated clusters in 3 dimensions.* Each cluster is generated as follows: set $x_1 = x_2 = x_3 = t$ with t taking on 100 equally spaced valued from -0.5 to 0.5 and then Gaussian noise with standard deviation .1 is added to each feature. Cluster 2 is generated in the same way, except that the value 10 is then added to each feature. The result is two elongated clusters, stretching out along the main diagonal of a three dimensional cube.

Fifty realizations were generated from each setting. In the non-null settings, the clusters have no overlap, so that there is no confusion over the definition of the "true" number of clusters.

In Tibshirani et al. (2001) a number of different methods for assessing the number of clusters were compared, and the Gap test performed best. Here

we enter the prediction strength estimate into the comparison: we select the number of clusters to be the largest k such that the $\mathrm{ps}(k) + \mathrm{se}(k) \geq .80$, where $\mathrm{se}(k)$ is the standard error of the prediction strength over the 5 cross-validation folds. [threshold values in the range .8 to .9 gave identical results.] For brevity, we compare this to the Gap test, and the methods due to Calinski & Harabasz (1974) and Krzanowski & Lai (1985). The first method uses

$$CH(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)} \tag{6}$$

where $B(k)$ and $W(k)$ are the between and within cluster sums of squares, with k clusters. $CH(k)$ is maximized over the number of clusters k. $CH(1)$ is not defined; even if it were modified by replacing $k-1$ with k, its value at 1 would be zero. Since $CH(k) > 0$ for $k > 1$, the maximum would never occur at $k = 1$. Krzanowski & Lai (1985) define

$$\mathrm{DIFF(k)} = (k-1)^{2/p} W_{k-1} - k^{2/p} W_k \tag{7}$$

and choose k to maximize the quantity

$$\mathrm{KL}(k) = \Big| \frac{\mathrm{DIFF(k)}}{\mathrm{DIFF(k+1)}} \Big|. \tag{8}$$

The results of the simulation study are given in Table 2. The prediction strength estimate does well in all but the last scenario, on a par with the gap statistic. In the last scenario, the clusters are long and narrow, and use of the principal component parameterization dramatically improves the gap test (see Tibshirani et al. (2001)).

7 Asymptotic properties of prediction strength

Here we give a theoretical justification for prediction strength, in the context of the k-means clustering algorithm. We consider k_0 populations that are given by uniform distributions on k_0 unit balls in d-space ($d > 1$), whose centers have pairwise distances of at least 4. Considering such well-separated simple clusters allows to clearly present the main arguments without obscuring them with lengthy technicalities. The following result shows that $\mathrm{ps}(k)$ exhibits indeed a sharp drop from 1 at k_0:

Table 2: *Results of simulation study. Numbers are counts out of 50 trials. "*" indicates column corresponding to correct number of clusters.*

Method	Estimate of number of clusters $\hat{k}$ 1	2	3	4	5	6	7	8	9	10
	Null model in 10 dimensions									
Gap/unif	49*	1	0	0	0	0	0	0	0	0
CH	0*	50	0	0	0	0	0	0	0	0
KL	0*	29	5	3	3	2	2	0	0	0
Pred str	50*	0	0	0	0	0	0	0	0	0
	Three cluster model									
Gap/unif	1	0	49*	0	0	0	0	0	0	0
CH	0	0	50*	0	0	0	0	0	0	0
KL	0	0	39*	0	5	1	1	2	0	0
Pred str	0	0	49*	1	0	0	0	0	0	0
	Random 4 cluster model in 3 dims.									
Gap/unif	0	1	2	47*	0	0	0	0	0	0
CH	0	0	0	42*	8	0	0	0	0	0
KL	0	0	0	35*	5	3	3	3	0	0
Pred str	0	0	0	50*	0	0	0	0	0	0
	Random 4 cluster model in 10 dims.									
Gap/unif	0	0	0	50*	0	0	0	0	0	0
CH	0	1	4	44*	1	0	0	0	0	0
KL	0	0	0	45*	3	1	1	0	0	0
Pred str	0	0	0	49*	1	0	0	0	0	0
	Two elongated clusters									
Gap/unif	0	0*	17	16	2	14	1	0	0	0
CH	0	0*	0	0	0	0	0	7	16	27
KL	0	50*	0	0	0	0	0	0	0	0
Pred str	0	27*	2	19	0	0	0	0	0	0

Theorem 1

$$\begin{aligned} \mathrm{ps}(k_0) &= 1 + o_p(1) \\ \sup_{k_0+1 \le k \le M} \mathrm{ps}(k) &\le \frac{2}{3} + o_p(1) \end{aligned}$$

Thus $\hat{k}$ is consistent for estimating k_0.

The dependence of $\mathrm{ps}(k)$ on the sample size n is suppressed in the notation. Also, it is possible to extend the theorem to let M increase with n.

Proof: Denote the k_0 population means by $m_1, \ldots, m_{k_0}$, and the k_0 optimal k-means centroids for the training and test sets by $\{\hat{m}_i^{tr}\}$ and $\{\hat{m}_i^{te}\}$, respectively. Theorem 3 in Pollard (1982) with a simple modification (see the example following said theorem) implies that for an appropriate labeling of the centroids

$$\sup_{1 \le i \le k_0} |\hat{m}_i^{tr} - m_i| = o_p(1), \quad \sup_{1 \le i \le k_0} |\hat{m}_i^{te} - m_i| = o_p(1). \tag{9}$$

But as soon as the above suprema are small enough (under the assumptions made for this theorem it is enough if the sup are smaller than 1), then all test data from the ith population ($1 \le i \le k_0$) are assigned to a common training centroid and to a common test centroid. But then $\mathrm{ps}(k_0) = 1$. Together with (9) this shows $\mathrm{ps}(k_0) = 1 + o_p(1)$.

Next let $k > k_0$. Considerations similar to those leading to (9) show that for n large enough, one of the k_0 populations, say the first, will have two test data centroids. For simplicity we consider only the case where there are exactly two such centroids. Then the test data falling into the support $B(m_1)$ of the first population are split into two clusters by the boundary of a halfspace H_{te}. Likewise, one population is split into two clusters by a halfspace H_{tr} from the training data clustering. We consider now the important case where the splits of the training and test clustering occur in the same population. The other cases are dealt with similarly.

From the definition of $\mathrm{ps}(k)$,

$$\mathrm{ps}(k) \le \text{cv-ave} \frac{1}{n_{k1}(n_{k1}-1)} \sum_{i \ne j \in A_{k1}} 1\Big(D[C(X_{tr}, k), X_{te}]_{ij} = 1\Big)$$

$$= \text{ cv-ave } \frac{(n/2)^2}{\sum_{1\le i\ne j\le n/2} 1\Big(\text{both } \underline{X}_{te,i} \text{ and } \underline{X}_{te,j} \text{ fall into } B(m_1)\cap H_{te}\Big)}$$
$$\times \quad \frac{\sum_{1\le i\ne j\le n/2} 1\Big(\text{both } \underline{X}_{te,i} \text{ and } \underline{X}_{te,j} \text{ fall into } B(m_1)\cap H_{te}\cap H_{tr} \text{ or}}{(n/2)^2}$$
$$\text{or } B(m_1)\cap H_{te}\cap H_{tr}^c\Big) \qquad (10)$$

The random halfspaces H_{te} and H_{tr} are independent; by a symmetry argument, their normal directions are distributed uniformly on the unit sphere S^{d-1}, and the distance of the bounding hyperplane to m_1 converges to zero. By the uniform strong law for U-statistics (see Thm. 7 in (Nolan & Pollard 1997)), $\frac{1}{(n/2)^2}\sum_{1\le i\ne j\le n/2} 1\Big(\text{both } \underline{X}_{te,i} \text{ and } \underline{X}_{te,j} \text{ fall into } B(m_1)\cap H\Big)$ converges a.s. to $P^2(\underline{X}_{te,1}\in B(m_1)\cap H)$ uniformly over all halfspaces $H\subset \mathbf{R}^d$. Hence (10) equals

$$\frac{E\,P^2(\underline{X}_{te,1}\in B(m_1)\cap H_1\cap H_2|H_1,H_2)+E\,P^2(\underline{X}_{te,1}\in B(m_1)\cap H_1\cap H_2^c|H_1,H_2)}{E\,P^2(\underline{X}_{te,1}\in B(m_1)\cap H_1|H_1)}+o_p(1) \qquad (11)$$

as both n and the number of cross-validation splits becomes large. Here H_1 and H_2 are halfspaces whose bounding hyperplanes contain m_1 and whose normal vectors are independently distributed on S^{d-1}.

Clearly $P(\underline{X}_{te,1}\in B(m_1)\cap H_1|H_1)=\frac{1}{2k_0}$. Further $P(\underline{X}_{te,1}\in B(m_1)\cap H_1\cap H_2|H_1,H_2)=\frac{1-\theta/\pi}{2k_0}$, where $\theta\in(0,\pi)$ is the angle between the normals of H_1 and H_2, and $P(\underline{X}_{te,1}\in B(m_1)\cap H_1\cap H_2^c|H_1,H_2)=\frac{\theta/\pi}{2k_0}$. It follows from formula (2.2.7) Watson (1983) that said angle θ has density $g(\theta)=\frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d-1}{2})\sqrt{\pi}}(\sin\theta)^{d-2}$. Hence the numerator in (11) equals

$$\frac{1}{4k_0^2}\int_0^\pi (1-\theta/\pi)^2 g(\theta)d\theta+\frac{1}{4k_0^2}\int_0^\pi(\theta/\pi)^2 g(\theta)d\theta=\frac{1}{4k_0^2}\int_0^\pi p(\theta)g(\theta)d\theta, \quad (12)$$

where $p(\theta):=(1-\theta/\pi)^2+(\theta/\pi)^2$ is symmetric around $\theta=\pi/2$ and strictly decreasing on $(0,\pi/2)$. Thus there is a $\bar\theta\in(0,\pi/2)$ such that $\bar p(\theta):=p(\theta)-\frac{1}{\pi}\int_0^\pi p(\theta)d\theta$ is negative in $(\bar\theta,\pi-\bar\theta)$ and positive outside this interval. $\bar g(\theta):=g(\theta)-g(\bar\theta)$ is positive in $(\bar\theta,\pi-\bar\theta)$ and negative outside this interval, by symmetry. So $\int_0^\pi \bar p(\theta)\bar g(\theta)d\theta\le 0$ and hence (12) equals

$$\frac{1}{4k_0^2}\Big(\int_0^\pi \bar p(\theta)\bar g(\theta)d\theta+g(\bar\theta)\int_0^\pi \bar p(\theta)d\theta+\frac{1}{\pi}\int_0^\pi p(\theta)d\theta\int_0^\pi g(\theta)d\theta\Big)$$

$$
\begin{aligned}
&\leq \frac{1}{4k_0^2\pi}\int_0^\pi p(\theta)d\theta \qquad\qquad \text{as } \int_0^\pi \bar{p}(\theta)d\theta = 0,\ \int_0^\pi g(\theta)d\theta = 1 \\
&= \frac{1}{2k_0^2\pi}\int_0^\pi (\theta/\pi)^2 d\theta \\
&= \frac{1}{6k_0^2}
\end{aligned}
$$

Thus (11) is not larger than $\frac{2}{3} + o_p(1)$. It follows from the above arguments that this bound is uniform over $k \in \{k_0 + 1, \ldots, M\}$, where M can also be allowed to grow appropriately with n. $\square$.

Acknowledgments: We would like to thank Gil Chu and Trevor Hastie for helpful discussions. Tibshirani was partially supported by NIH grant 2 R01 CA72028, and NSF grant DMS-9971405. Walther was partially supported by grants DMS-9704557 and DMS-9875598.

References

Calinski, R. B. & Harabasz, J. (1974), 'A dendrite method for cluster analysis', *Communications in statistics* **3**, 1–27.

Gordon, A. (1999), *Classification (2nd edition)*, Chapman and Hall/CRC press, London.

Krzanowski, W. J. & Lai, Y. T. (1985), 'A criterion for determining the number of groups in a data set using sum of squares clustering', *Biometrics* **44**, 23–34.

Milligan, G. W. & Cooper, M. C. (1985), 'An examination of procedures for determining the number of clusters in a data set', *Psychometrika* **50**, 159–179.

Nolan, D. & Pollard, D. (1997), 'U-processes: Rates of convergence', *Ann. Stat.* **15**, 780–799.

Perou, C., Jeffrey, S., van de Rijn, M., Rees, C., Eisen, M., Ross, D., Pergamenschikov, A., Williams, C., Zhu, S., Lee, J., Lashkari, D., Shalon, D., Brown, P. & Botstein, D. (1999), 'Distinctive gene expression patterns in human mammary epiphelial cells and breast cancers', *Proc .Nat. Acad. Sci.* **96**, 9212–9217.

Pollard, D. (1982), 'A central limit theorem for k-means clustering.', *Ann. Prob.* **19**, 919–926.

Sugar, C. (1998), Techniques for clustering and classification with applications to medical problems, Technical report, Stanford University. Ph.D. dissertation in Statistics, R. Olshen supervisor.

Sugar, C., Lenert, L. & Olshen, R. (1999), An application of cluster analysis to health services research: Empirically defined health states for depression from the sf-12., Technical report, Stanford University.

Tibshirani, R., Walther, G. & Hastie, T. (2001), 'Estimating the number of clusters in a dataset via the gap statistic', *J. Royal. Statist. Soc. B. (to appear)* .

Watson, G. (1983), *Statistics on Spheres*, Wiley, New York.