Cluster Validation by Prediction Strength

Robert TIBSHIRANI and Guenther WALTHER

This article proposes 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 unlabeled 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.

Key Words: Number of clusters; Prediction; Unsupervised learning.

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 problem focus on the 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 were summarized in the comprehensive survey by Milligan and Cooper (1985), and Gordon (1999) discussed the best performers. More recent proposals include Tibshirani, Walther, and Hastie (2001), Sugar (1998), and Sugar, Lenert, and Olshen (1999). It is not clear, however, if these methods are widely used; this may be because they are difficult to intrepret.

In this article we take a different approach. We view estimation of the number of

Robert Tibshirani is Professor, Department of Health Research and Policy and Department of Statistics, Stanford University, Stanford, CA 94305 (E-mail: tibs@stat.stanford.edu). Guenther Walther is Associate Professor, Department of Statistics, Stanford University, Stanford, CA 94305 (E-mail: walther@stat.stanford.edu).

Journal of Computational and Graphical Statistics, Volume 14, Number 3, Pages 511–528
DOI: 10.1198/106186005X59243

clusters as a model selection problem. In classification problems with labeled 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 focusing on prediction error rather than the within-cluster 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.

There are several recent proposals for inference on clustering in a high-dimensional setting, such as microarrays. Kerr and Churchill (2001) used an analysis of variance model to estimate differential expression of genes across multiple conditions and to account for other sources of variation in microarray data. Residuals from the fitted ANOVA model provide an estimate of the error distribution, which allows us to resample gene expression and thus obtain a number of bootstrap clusterings. Clusters are obtained by correlating genes to one out of a number of temporal profiles. A match of a gene to a profile is declared 95% stable if it occurs in the analysis of the original data and in at least 95% of the bootstrap clusterings.

Yeung, Haynor, and Ruzzo (2001) provided a framework for comparing different clustering algorithms for gene expression data. They defined a "Figure of Merit" for assessing the predictive power of an algorithm by leaving out one experimental condition in turn, clustering genes based on the remaining data, and then measuring the within-cluster similarity of expression values in the experimental condition left out. The Figures of Merit are plotted over a range of different numbers of clusters. Typically, some clustering algorithms exhibit Figure of Merit curves that dominate other algorithms, yielding a criterion for choosing the most appropriate algorithm. It is not clear, however, if and how this methodology can be extended for inference on the number of clusters.

If a parametric model for the cluster components is appropriate, such as the normal model, then model-based clustering allows inference about the number of clusters; see, for example, Fraley and Raftery (1998). Model-based clustering employs the EM algorithm to estimate the parameters in a Gaussian mixture, where the covariance matrix is appropriately restricted to keep the number of parameters manageable. The Bayesian information criterion (BIC) can be used to select the model, that is, the number of components in the mixture. An important advantage of this model-based approach is that it allows principled inference about various quantities, such as the number of clusters or the uncertainty in classifying individual observations. Yeung et al. (2001) successfully applied model-based clustering to gene expression data.

Ben-Hur, Elisseeff, and Guyon (2002) proposed a stability-based criterion for determining the number of clusters. One hundred pairs of subsamples of the data are generated, and each subsample is clustered into k clusters with average-link hierarchical clustering. For each of the 100 pairs, the observations contained in both subsamples are extracted, and a similarity measure (such as the Jaccard coefficient) is computed for their two clustering outcomes. The similarity measure takes values between 0 and 1, with a value close to 1 indicating that the two clusterings are the same for most observations in the joint subsample. The histogram for the 100 similarity measures is plotted, and the process is repeated for a range of clusters k. The estimated number of clusters is then taken to be that value of k where a transition occurs from similarity values concentrated near 1 to a distribution with wider spread below 1. The idea of examining the stability of clusters is similar to the prediction strength idea presented here. However, it was pointed out by Ben-Hur, Elisseeff, and Guyon (2002) that a trial value k smaller than the true value k_o can readily lead to a premature spread in the distribution of the similarity measure, resulting in an underestimate of the number of clusters. Likewise, if the procedure is applied to the trial value $k_o + 1$, then we expect that one cluster will be split into two, while the clustering outcomes for the other $k_o - 1$ clusters will be same. Hence the similarity measure will be about $(k_o - 1)/k_o$, which for $k_o > 9$ is close enough to 1 to result in an overestimate of the number of clusters. Indeed, no theoretical justification of that methodology was provided by Ben-Hur, Elisseeff, and Guyon (2002). The prediction strength idea introduced here employs a quite different criterion, and we are able to provide a theoretical justification as well as a formal connection to the theory of supervised learning.

Section 2 describes the basic procedure for estimating prediction strength. Section 3 gives 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. Section 5 discusses 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, Section 7 establishes consistency of the method in a simple but informative case.

2. PREDICTION STRENGTH OF CLUSTERING

Our training data $X_{\text{tr}} = \{x_{ij}\}, i = 1, 2, \ldots n;, j = 1, 2, \ldots p$ 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 $D[C(\ldots), X_{\text{tr}}]$ be an $n \times n$ matrix, with ii'th element $D[C(\ldots), X_{\text{tr}}]_{ii'} = 1$ if observations i and i' fall into the same cluster, and zero otherwise. We call these entries "co-memberships." In general, the clustering $C(\ldots)$ need not be derived from X_{tr}. For example, we can apply the k-means algorithm to some dataset Y, which will result in a partition of the observation space into k polygonal regions. If we denote this clustering by $C(Y, k)$, then $D[C(Y, k), X_{\text{tr}}]_{ii'} = 1$ if observations i and i' of X_{tr} fall into the same polygonal region of $C(Y, k)$.

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

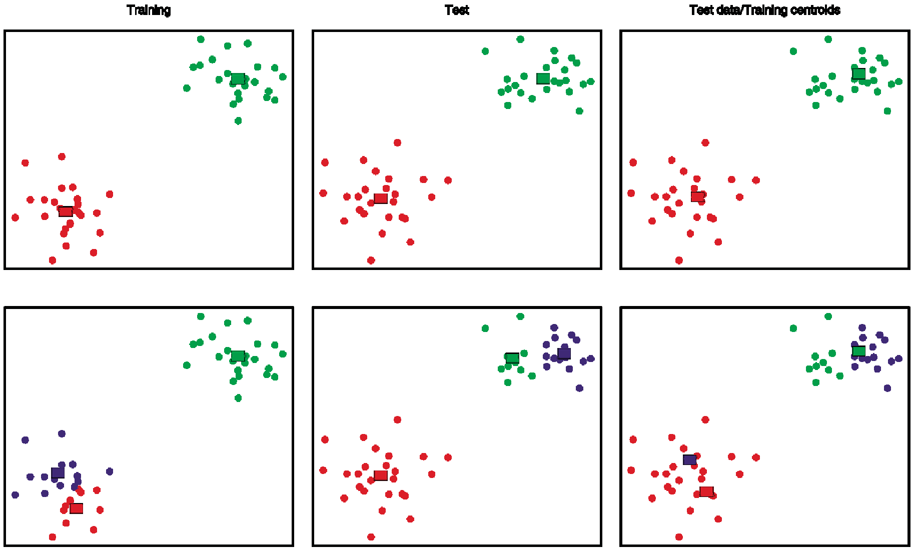

Figure 1. Illustration of prediction strength idea. Data are simulated in two well-separated clusters. In the top row k-means clustering with two centroids is applied to both the training and test data. In the top right panel, the training centroids classify the test points into the same two green and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

X_{te} into k clusters via an operation $C(X_{\mathrm{te}}, k)$, and summarize the cluster co-memberships via the $m \times m$ matrix $D[C(X_{\mathrm{te}}, k), X_{\mathrm{te}}]$.

The main idea of this article is to (1) cluster the test data into k clusters; (2) cluster the training data into k clusters, and then (3) 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 this idea. The data lie in two clusters. In the top row two-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 $A_{k1}, A_{k2}, \dots A_{kk}$ be the indices of the test observations in test clusters $1, 2, \dots k$. Let $n_{k1}, n_{k2}, \dots n_{kk}$ be the number of observations in these clusters.

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

$$\mathrm{ps}(k) = \min_{1 \le j \le k} \frac{1}{n_{kj}(n_{kj}-1)} \sum_{i \ne i' \in A_{kj}} D[C(X_{\mathrm{tr}}, k), X_{\mathrm{te}}]_{ii'}. \tag{2.1}$$

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 $\mathrm{ps}(k)$ will be high. Note that $\mathrm{ps}(1) = 1$ in general, because 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 $\mathrm{ps}(k)$ to be much smaller. Using the minimum rather than the average in expression (2.1) 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 focusing only on the pairwise co-memberships in (2.1), we finesse this problem. The identity of the cluster

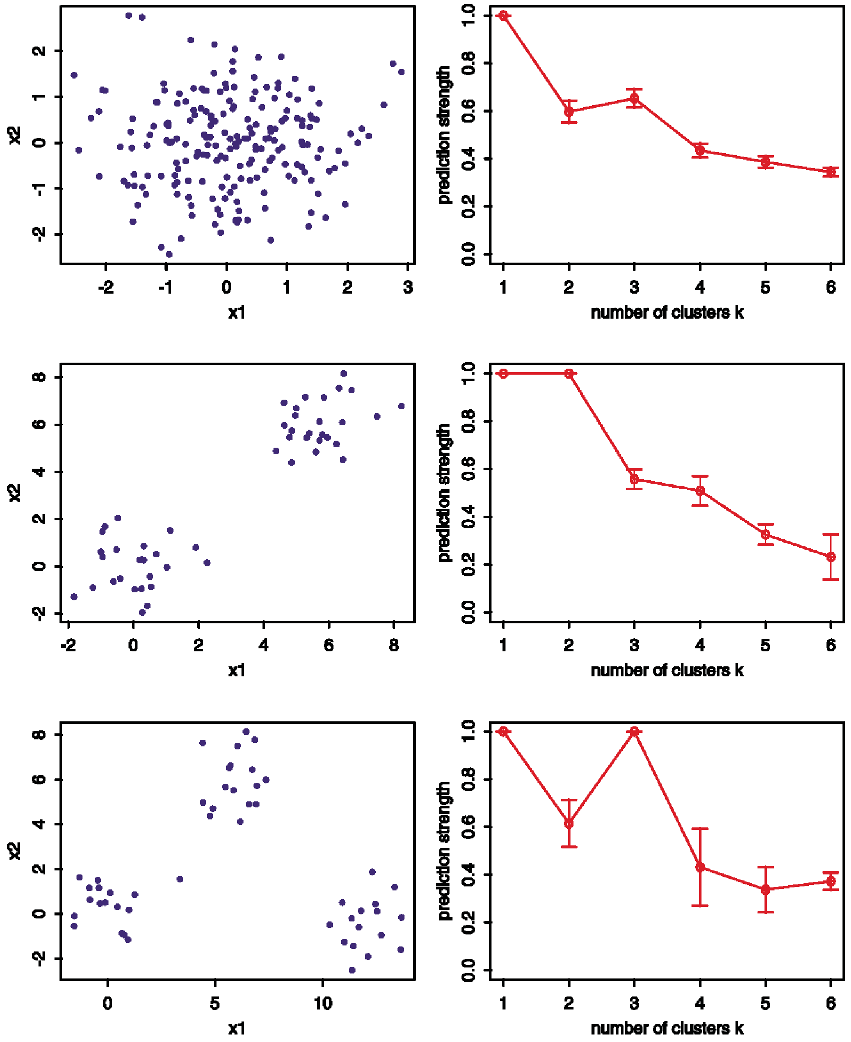

Figure 2. Results for one-, two-, and three-cluster examples. The test data are on the left, and prediction strength on the right. The vertical bars on the right give the standard error of the prediction strength over five cross-validation folds.

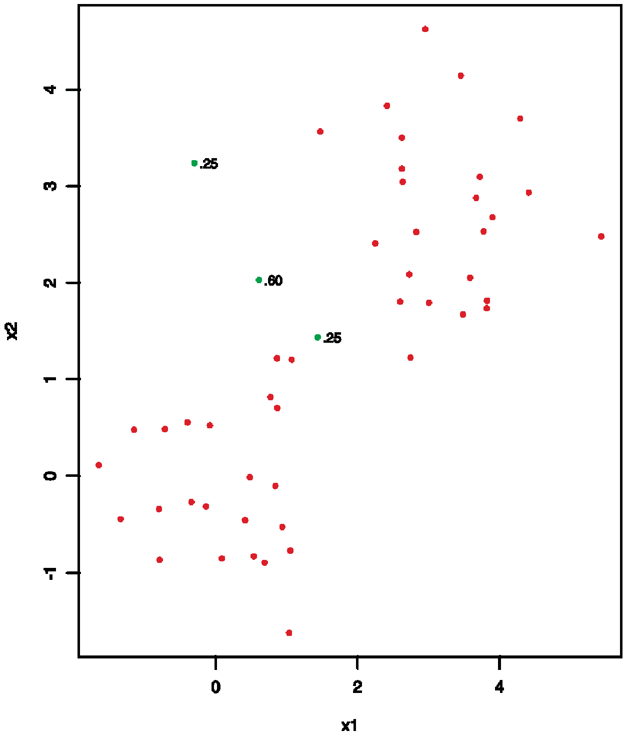

Figure 3. Individual prediction strengths, when the data shown are clustered into two clusters. Green: ps $< .90$ (prediction strength indicated); Red: ps $> .9$. We used the test sample shown, and five randomly generated training samples from the same population. The predictions strengths were estimated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

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 article show that a threshold in the range .8–.9 works for well separated clusters. We think of $\hat{k}$ as the largest number of clusters that can be reliably 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 (2.1). 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 two-fold and five-fold cross-validation. Their performance is quite similar, and we settle on two-fold cross-validation for the rest of the article.

Prediction strengths for individual observations can also be defined. Specifically, we define the prediction strength for observation i as

$$\text{ps}(i,k) = \frac{1}{\#A_k(i)} \cdot \sum_{i' \in A_k(i)} 1\left(D[C(X_{\text{tr}},k),X_{\text{te}}]_{ii'} = 1\right), \tag{2.2}$$

where $A_k(i)$ are the observations indices i' such that $i \neq i'$ and $D[C(X_{\text{te}}, k), X_{\text{te}}]_{ii'} = 1$. 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

This section provides 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, that is, $D[C^*(X), X]_{ij} = 1$ iff $\underline{X}_i$ and $\underline{X}_j$ are from the same group. Define the prediction error (loss) of the clustering procedure C by

$$\text{err}_C(k) = \frac{1}{n^2} \sum_{i,j=1}^{n} \left| D[C^*(X), X]_{ij} - D[C(X,k), X]_{ij} \right|. \tag{3.1}$$

As all matrix entries are either 0 or 1, one sees that $\text{err}_C(k)$ decomposes into two parts:

$$\text{err}_C(k) = \begin{pmatrix} \text{proportion of pairs} \\ (\underline{X}_i, \underline{X}_j) \text{ that } C(X,k) \\ \text{erroneously assigns to} \\ \text{the same group} \end{pmatrix} + \begin{pmatrix} \text{proportion of pairs} \\ (\underline{X}_i, \underline{X}_j) \text{ that } C(X,k) \\ \text{erroneously assigns to} \\ \text{different groups} \end{pmatrix}. \tag{3.2}$$

The first term tends to decrease as k increases (as fewer groups are erroneously aggregated into one big group), and 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. We can try to mimic this decomposition to estimate k, by letting $C(X_{\text{te}}, k)$ and $C(X_{\text{tr}}, k)$ 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 $C(X_{\text{te}}, k)$ 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

$$\max_{1 \leq j \leq k} \frac{1}{n_{kj}(n_{kj} - 1)} \sum_{i \neq i' \in A_{kj}} 1\left(D[C(X_{\text{tr}}, k), X_{\text{te}}]_{ii'} = 0 \right), \tag{3.3}$$

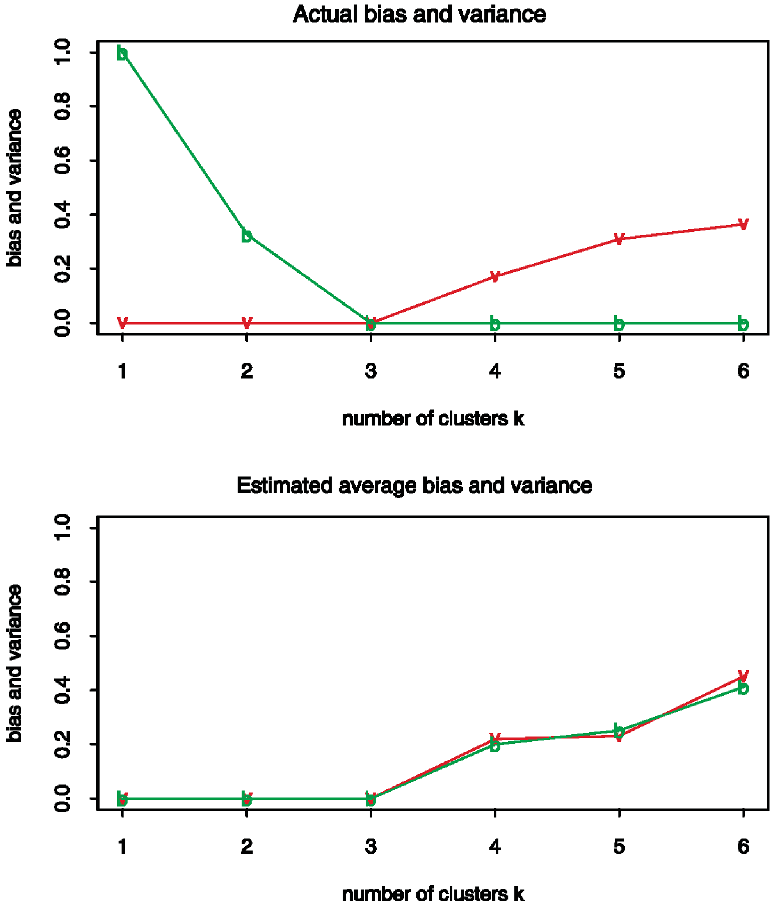

Figure 4. Bias and variance for the three-cluster example in Figure 2. The top panel shows the actual bias and variance using the real class labels, as in (3.2). The bottom panel uses $C(X_{te}, k)$ in place of the true labels $C^*(X)$.

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

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

where we modified the preliminary definition (2.1) 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 $\text{ps}(k) \geq .8$ or .9 works well in practice. We think of $\hat{k}$ as the largest number of clusters that can be accurately predicted in the dataset.

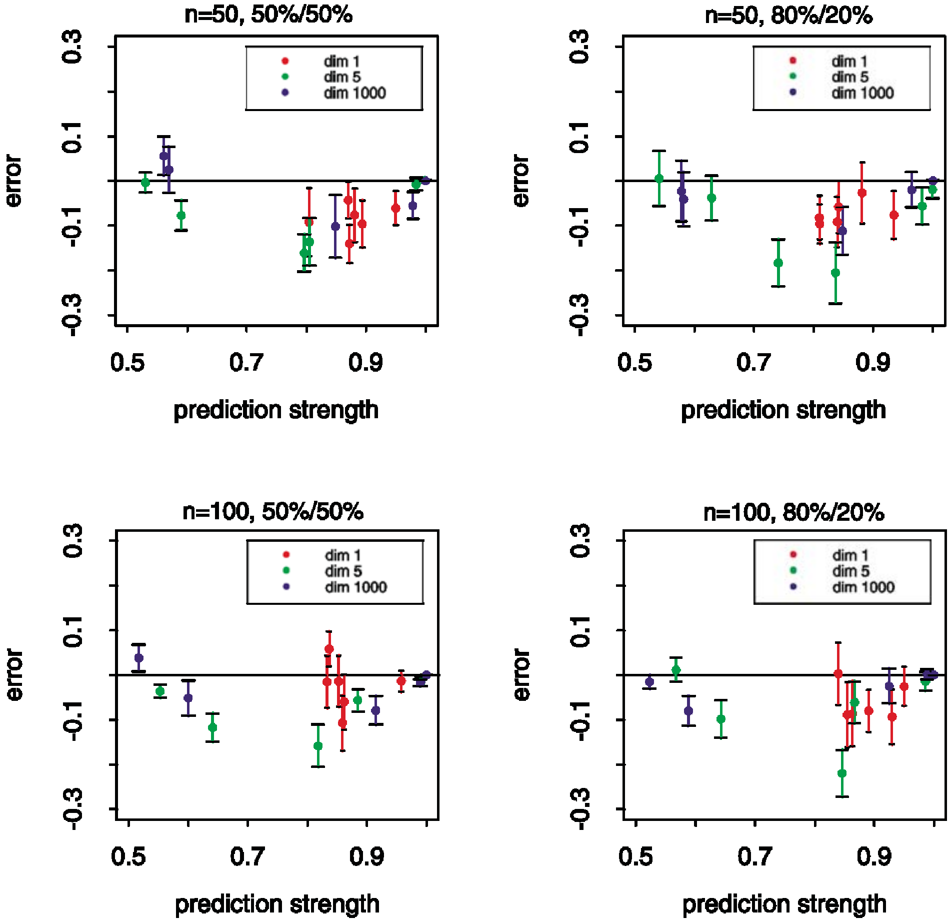

Figure 5. Bias in prediction strength estimates, for the experiments described in the text. The left panel shows the error $ps_{n/2,n/2} - ps_{n,\infty}$ *plotted as a function of the "true" prediction strength* $ps_{n,\infty}$*. The right panel assesses five-fold cross-validation, and hence* $ps_{4n/5,n/5}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

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 five-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 $\mathrm{ps}_{n/2,n/2} - \mathrm{ps}_{n,\infty}$ plotted as a function of the "true" prediction strength $\mathrm{ps}_{n,\infty}$. In the right panel we assess five-fold cross-validation, and hence we have plotted $\mathrm{ps}_{4n/5,n/5} - \mathrm{ps}_{n,\infty}$. In each case we show the mean over 10

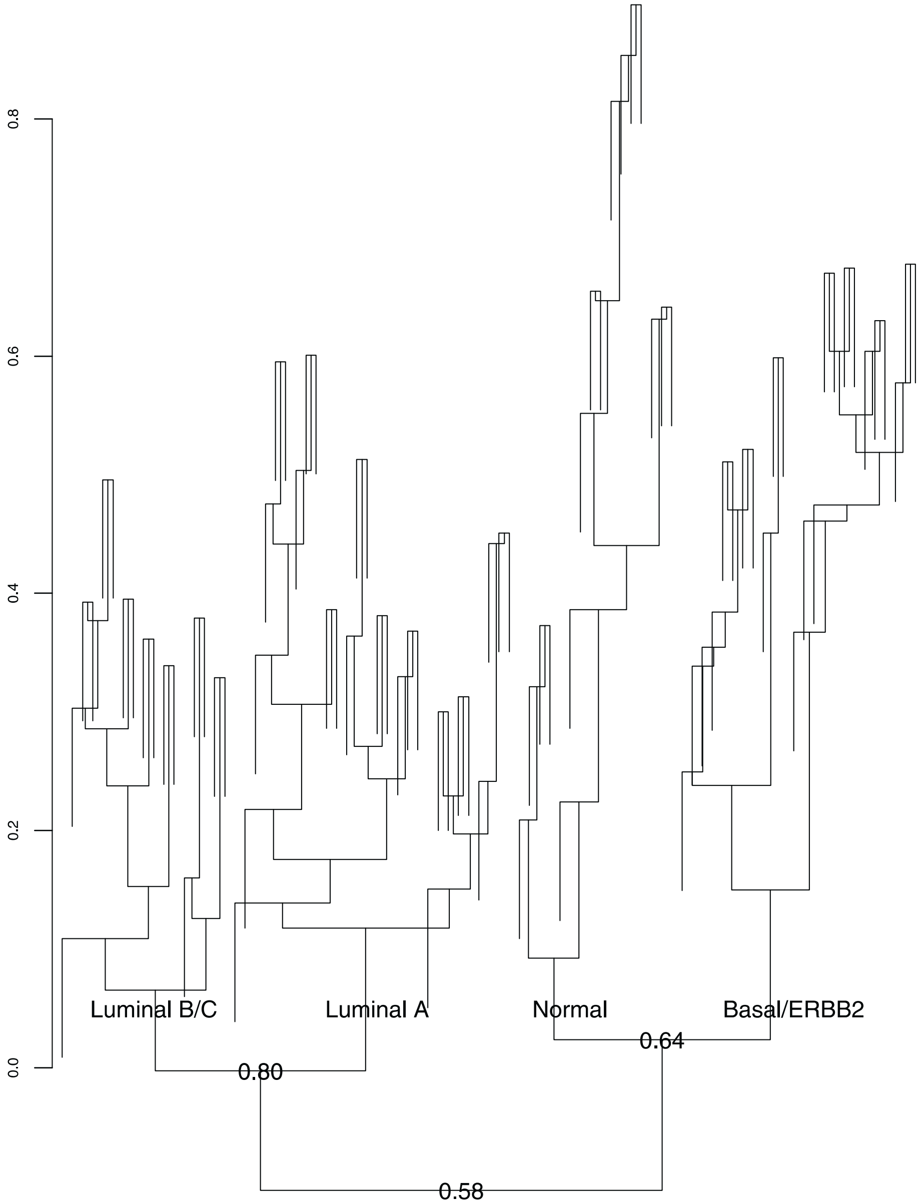

Figure 6. Dendrogram from breast cancer study, with the estimated prediction strength at the upper branches.

simulations, with one standard-error bands. There appears to be no advantage in using five-fold over two-fold cross-validation, and hence we used the latter in this article.

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. Clustering DNA microarray data is considered a hard problem not only because of the large dimension of the data, but also because there may be no underlying "true" number of clusters; the expression levels of some genes may not vary consistently with other genes, and clusters may have varying widths.

Figure 6 shows a dendrogram from hierarchical clustering of the gene expression of 85 breast cancer patients. These data are 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, labeled in the dendrogram. [This example is for illustration purposes. Our dendrogram is not the same as that of Perou et al. (1999). They used a nonstandard 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.] Hierarchical clustering is preferred to K-means 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 $C(X_{\text{tr}}, k)$, cutting off the resulting dendrogram at a height that produces k clusters. We tried this, and it produced prediction strengths $(1.00, .53, .42, .34)$ for $k = 1, 2, \ldots 4$. Thus we would conclude that none of the clusters is significant. However, this may not be the best strategy, as hierarchical clustering is performed bottom-up, and the resulting groups might look nothing like the original ones. As an alternative, we tried the following strategy: use hierarchical clustering to find potential clusters as in Figure 6, but then use the k-means clustering as $C(X_{\text{tr}}, k)$ 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 labeled 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 .59. We can look deeper by computing the prediction strength of all pairs of the four 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.

Table 1. Prediction Strength for All Pairs of the Four 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

6. A SIMULATION STUDY

In this section we replicate the simulation study done by Tibshirani, Walther, and Hastie (2001), comparing a number of different methods for estimating the number of clusters. We now include the prediction strength method in the comparison. We also add three difficult cluster scenarios to show the limits of the prediction strength methodology as well as its performance in high-dimensional settings, such as microarray analyses.

We thus generated datasets in eight different scenarios:

1. *Null (single cluster) data in 10 dimensions:* 200 data points uniformly distributed over the unit square in 10 dimensions.

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

3. *Four clusters in three 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. *Four 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. *Four clusters in two dimensions that are not well separated.* each cluster has 25 standard normal observations, centered at (0, 0), (0, 2.5), (2.5, 0) and (2.5, 2.5).

6. *Two elongated clusters in three dimensions.* Each cluster is generated as follows: set $x_1 = x_2 = x_3 = t$ with t taking on 100 equally spaced values from $-.5$ to .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.

7. *Two close and elongated clusters in three dimensions.* As in the previous scenario, with cluster 2 being generated in the same way as cluster 1, except that the value 1 is then added to the first feature only.

8. *Three clusters in a microarray-like setting.* Each of the three clusters has 33 standard normal observations in 1,000 dimensions, with each of the first 100 coordinates shifted by -2, 0, and 2, respectively.

Fifty realizations were generated from each setting.

In Tibshirani, Walther, and Hastie (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 $\text{ps}(k) + \text{se}(k) \geq .80$, where $\text{se}(k)$ is the standard error of the prediction strength over the five cross-validation folds. (Threshold values in the range .8 to .9 gave identical results.) We compare two applications of the prediction strength, one to k-kmeans ("Pred str") and one to hierarchical clustering ("Pred str/hc"), to the Gap test with uniform reference distribution ("Gap/unif") and principal component parameterization ("Gap/pc"), and to the methods due to Calinski and Harabasz (1974) and Krzanowski and Lai (1985). The first method uses

$$\text{CH}(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)}, \tag{6.1}$$

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

$$\text{DIFF}(k) = (k-1)^{2/p} W_{k-1} - k^{2/p} W_k, \tag{6.2}$$

and chose k to maximize the quantity

$$\text{KL}(k) = \left| \frac{\text{DIFF}(k)}{\text{DIFF}(k+1)} \right|. \tag{6.3}$$

The results of the simulation study are given in Table 2. The Gap/pc method was not evaluated on the microarray example, as it is not clear how to apply it with the number of variables larger than the sample size.

6.1 Discussion of the Simulation Results

The prediction strength estimate does well compared to the other methods in all scenarios except for the elongated clusters in scenario 6, where the KL and Gap/pc methods are the best performers. In that scenario, the clusters are long and narrow, and use of the principal component parameterization dramatically improves the Gap test (see Tibshirani, Walther, and Hastie 2001). The not-well-separated clusters in scenarios 5 and 7 proved too challenging for all methods; the CH criterion has the largest number of correct results in scenario 5, but this corresponds to a success rate of only 26%, with a large number of simulations resulting in a considerable overestimate.

Overall, the simulation study shows that the prediction strength estimate compares well to the other methods except for strongly elongated clusters. Also, in that scenario the

Table 2. Results of Simulation Study. Numbers are counts out of 50 trials. Counts for estimates larger than 10 are not displayed. "*" indicates column corresponding to correct number of clusters.

	Estimate of number of clusters $\hat{k}$									
Method	*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
Gap/pc	50*	0	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
Pred str/hc	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
Gap/pc	2	0	48*	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
Pred str/hc	0	0	46*	4	0	0	0	0	0	0
	Random four-cluster model in three dimensions									
Gap/unif	0	1	2	47*	0	0	0	0	0	0
Gap/pc	2	2	4	42*	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
Pred str/hc	0	0	0	34*	16	0	0	0	0	0
	Random four-cluster model in 10 dimensions									
Gap/unif	0	0	0	50*	0	0	0	0	0	0
Gap/pc	0	0	4	46*	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
Pred str/hc	0	0	0	31*	13	4	0	0	0	0

prediction strength performs better when applied to hierarchical clustering rather than k-means. This can be explained by the fact that k-means is implicitly biased towards spherical clusters, so when the prediction strength is applied to k-means, it selects the best model for the data from among models consisting of unions of spheres. As one referee pointed out, an elongated structure is perhaps best modeled as a union of spheres, so this approach is not inappropriate. Thus, although the choice of the clustering algorithm clearly needs to take into account the structure of the clusters, the prediction strength has proven effective in selecting an appropriate model from those under consideration.

7. ASYMPTOTIC PROPERTIES OF PREDICTION STRENGTH

This section gives 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 four. Considering such well-separated simple clusters allows us to clearly present the

Table 2. Continued.

	Estimate of number of clusters $\hat{k}$									
Method	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
	Four not well-separated clusters in two dimensions									
Gap/unif	50	0	0	0*	0	0	0	0	0	0
Gap/pc	49	1	0	0*	0	0	0	0	0	0
CH	0	0	1	13*	7	4	4	2	8	11
KL	0	13	4	7*	4	2	9	5	0	0
Pred str	35	12	2	1*	0	0	0	0	0	0
Pred str/hc	48	2	0	0*	0	0	0	0	0	0
	Two elongated clusters									
Gap/unif	0	0*	17	16	2	14	1	0	0	0
Gap/pc	0	50*	0	0	0	0	0	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
Pred str/hc	0	42*	8	0	0	0	0	0	0	0
	Two close and elongated clusters									
Gap/unif	5	0*	0	0	7	32	6	0	0	0
Gap/pc	50	0*	0	0	0	0	0	0	0	0
CH	0	0*	0	0	0	27	8	14	1	0
KL	0	0*	0	19	1	12	8	6	0	0
Pred str	9	7*	1	31	0	2	0	0	0	0
Pred str/hc	44	6*	0	0	0	0	0	0	0	0
	Three clusters in microarray setting									
Gap/unif	30	12	8*	0	0	0	0	0	0	0
CH	0	50	0*	0	0	0	0	0	0	0
KL	0	24	26*	0	0	0	0	0	0	0
Pred str	0	0	50*	0	0	0	0	0	0	0
Pred str/hc	0	0	50*	0	0	0	0	0	0	0

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:

Theorem 1.

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

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 $\{\hat{m}_i^{\mathrm{tr}}\}$ and $\{\hat{m}_i^{\mathrm{te}}\}$, respectively. The theorem in Pollard (1982) with a simple modification (see the example following said theorem) implies that for an appropriate labeling of the centroids

$$\sup_{1 \leq i \leq k_0} |\hat{m}_i^{\mathrm{tr}} - m_i| = o_p(1), \quad \sup_{1 \leq i \leq k_0} |\hat{m}_i^{\mathrm{te}} - m_i| = o_p(1). \tag{7.1}$$

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 $\text{ps}(k_0) = 1$. Together with (7.1) this shows $\text{ps}(k_0) = 1 + o_p(1)$.

Next let $k > k_0$. Considerations similar to those leading to (7.1) 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 $\text{ps}(k)$,

$$\begin{aligned}
\text{ps}(k) \quad &\le \quad \text{cv-ave}\, \frac{1}{n_{k1}(n_{k1}-1)} \sum_{i \neq j \in A_{k1}} 1\left(D[C(X_{\text{tr}}, k), X_{\text{te}}]_{ij} = 1\right) \\
&= \quad \text{cv-ave}\, \frac{(n/2)^2}{\sum_{1 \le i \neq j \le n/2} 1\left(\text{both } \underline{X}_{\text{te},i} \text{ and } \underline{X}_{\text{te},j} \text{ fall into } B(m_1) \cap H_{\text{te}}\ \right)} \\
&\times \quad \frac{\sum_{1 \le i \neq j \le n/2} 1\left(\text{both } \underline{X}_{\text{te},i} \text{ and } \underline{X}_{\text{te},j} \text{ fall into } B(m_1) \cap H_{\text{te}} \cap H_{\text{tr}} \text{ or}\right.}{(n/2)^2} \\
& \qquad \left.\text{or } \ B(m_1) \cap H_{\text{te}} \cap H_{\text{tr}}^c\right). \qquad (7.2)
\end{aligned}$$

The random halfspaces H_{te} and H_{tr} are independent; by a symmetry argument, their normal directions are distributed uniformly on the unit sphere, and the distance of the bounding hyperplane to m_1 converges to zero. By the uniform strong law for U-statistics (see Nolan and Pollard 1997, theorem 7), $\frac{1}{(n/2)^2} \sum_{1 \le i \neq j \le n/2} 1\big(\text{both } \underline{X}_{\text{te},i}$ and $\underline{X}_{\text{te},j}$ fall into $B(m_1) \cap H\big)$ converges almost surely to $P^2(\underline{X}_{\text{te},1} \in B(m_1) \cap H)$ uniformly over all halfspaces $H \subset \mathbf{R}^d$. Hence (7.2) equals

$$\frac{E\, P^2(\underline{X}_{\text{te},1} \in B(m_1) \cap H_1 \cap H_2 | H_1, H_2)}{E\, P^2(\underline{X}_{\text{te},1} \in B(m_1) \cap H_1 | H_1)} + \frac{E\, P^2(\underline{X}_{\text{te},1} \in B(m_1) \cap H_1 \cap H_2^c | H_1, H_2)}{E\, P^2(\underline{X}_{\text{te},1} \in B(m_1) \cap H_1 | H_1)} + o_p(1) \qquad (7.3)$$

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 the unit sphere.

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

Hence the numerator in (7.3) equals

$$\frac{1}{4k_0^2}\int_0^\pi (1-\theta/\pi)^2 g(\theta)d\theta + \frac{1}{4k_0^2}\int_0^\pi (\theta/\pi)^2 g(\theta)d\theta = \frac{1}{4k_0^2}\int_0^\pi p(\theta)g(\theta)d\theta, \tag{7.4}$$

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 (7.4) equals

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

Thus, (7.3) 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. □

ACKNOWLEDGMENTS

We would like to thank the referees for providing the Ben-Hur et al., Kerr/Churchill, and Yeung et al. references.

[Received March 2002. Revised April 2004.]

REFERENCES

Ben-Hur, A., Elisseeff, A., and Guyon, I. (2002), "A Stability Based Method for Discovering Structure in Clustered Data," in *Proceedings of the Pacific Symposium on Biocomputing*, pp. 6–17.

Calinski, R. B., and Harabasz, J. (1974), "A Dendrite Method for Cluster Analysis," *Communications in Statistics*, 3, 1–27.

Fraley, C., and Raftery, A. (1998), "How Many Clusters? Which Clustering Method?—Answers via Model-Based Cluster Analysis," *Computer Journal*, 41, 578–588.

Gordon, A. (1999), *Classification* (2nd ed.), London: Chapman and Hall/CRC Press.

Kerr, M., and Churchill, G. (2001), "Bootstrapping Cluster Analysis: Assessing the Reliability of Conclusions from Microarray Experiments," in *Proceedings of the National Academy of Sciences*, pp. 8961–8965.

Krzanowski, W. J., and 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., and Cooper, M. C. (1985), "An Examination of Procedures for Determining the Number of Clusters in a Data set," *Psychometrika*, 50, 159–179.

Nolan, D., and Pollard, D. (1997), "U-Processes: Rates of Convergence," *The Annals of Statistics*, 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., and Botstein, D. (1999), "Distinctive Gene Expression Patterns in Human Mammary Epiphelial Cells and Breast Cancers," in *Proceedings of the National Academy of Sciences*, 96, 9212–9217.

Pollard, D. (1982), "A Central Limit Theorem for k-means Clustering," *Annals of Probability*, 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., and 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., and Hastie, T. (2001), "Estimating the Number of Clusters in a Dataset via the Gap Statistic," *Journal of the Royal Statistical Society*, Ser. B, 32, 411–423.

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

Yeung, K., Fraley, C., Murua, A., Raftery, A., and Ruzzo, W. (2001), "Model-Based Clustering and Data Transformations for Gene Expression Data," *Bioinformatics*, 17, 977–987.

Yeung, K., Haynor, D., and Ruzzo, W. (2001), "Validating Clustering for Gene Expression Data," *Bioinformatics*, 309–318.