Research

A prediction-based resampling method for estimating the number of clusters in a dataset

Sandrine Dudoit*‡ and Jane Fridlyand†‡

Addresses: *Division of Biostatistics, School of Public Health, University of California Berkeley, 140 Earl Warren Hall, Berkeley, CA 94720-7360, USA. †Jain Lab, Comprehensive Cancer Center, University of California San Francisco, 2340 Sutter St, San Francisco, CA 94143-0128, USA. ‡Both authors contributed equally to this work.

Correspondence: Sandrine Dudoit. E-mail: sandrine@stat.berkeley.edu

Published: 25 June 2002

Genome ***Biology*** 2002, **3(7)**:research0036.1–0036.21

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/7/research/0036

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Received: 18 February 2002
Revised: 22 April 2002
Accepted: 15 May 2002

Abstract

Background: Microarray technology is increasingly being applied in biological and medical research to address a wide range of problems, such as the classification of tumors. An important statistical problem associated with tumor classification is the identification of new tumor classes using gene-expression profiles. Two essential aspects of this clustering problem are: to estimate the number of clusters, if any, in a dataset; and to allocate tumor samples to these clusters, and assess the confidence of cluster assignments for individual samples. Here we address the first of these problems.

Results: We have developed a new prediction-based resampling method, Clest, to estimate the number of clusters in a dataset. The performance of the new and existing methods were compared using simulated data and gene-expression data from four recently published cancer microarray studies. Clest was generally found to be more accurate and robust than the six existing methods considered in the study.

Conclusions: Focusing on prediction accuracy in conjunction with resampling produces accurate and robust estimates of the number of clusters.

Background

The burgeoning field of genomics, and in particular DNA microarray experiments, has revived interest in cluster analysis by raising new methodological and computational challenges. DNA microarrays are part of a new and promising class of biotechnologies that allow the monitoring of expression levels in cells for thousands of genes simultaneously. Microarray experiments are increasingly being caried out in biological and medical research to address a wide range of problems, including the classification of tumors [1-6]. A reliable and precise classification of tumors is essential for successful diagnosis and treatment of cancer. By allowing the monitoring of expression levels on a genomic scale, microarray experiments may lead to a more complete understanding of the molecular variations among tumors and hence to a finer and more reliable classification. An important statistical problem associated with tumor classification is the identification of new tumor classes using gene-expression profiles. Two essential aspects of this clustering problem are: first, to accurately estimate the number of clusters, if any, in a dataset; and second, to allocate tumor samples accurately to these clusters, and assess the confidence of cluster assignments for individual samples. In a clinical application of microarray-based cancer diagnosis, the definition of new tumor classes would be based on the clustering results, and these classes would then be used to

build predictors for new tumor samples. Inaccurate cluster assignments could lead to erroneous diagnoses and unsuitable treatment protocols.

Here we address the estimation of the number of clusters in a dataset. First, we describe the basic principles of cluster analysis and review existing methods for estimating the number of clusters. We then present a new prediction-based resampling method, Clest, for estimating the number of clusters in a dataset. The performance of the new and existing methods is compared using simulated data and gene-expression data from four recently published cancer microarray studies. We have addressed the problem of improving and assessing the accuracy of a given clustering procedure in [7].

Cluster analysis

In classification, one is concerned with assigning objects to classes on the basis of measurements made on these objects. There are two main aspects to classification: discrimination and clustering, or supervised and unsupervised learning. In unsupervised learning (also known as cluster analysis, class discovery and unsupervised pattern recognition), the classes are unknown *a priori* and need to be discovered from the data. In contrast, in supervised learning (also known as discriminant analysis, class prediction, and supervised pattern recognition), the classes are predefined and the task is to understand the basis for the classification from a set of labeled objects (training or learning set). This information is then used to classify future observations. The present article focuses on the unsupervised problem, that is, on cluster analysis, but draws on notions from supervised learning to address the problem.

In cluster analysis, the data are assumed to be sampled from a mixture distribution with K components corresponding to the K clusters to be recovered. Let $(X_1, \ldots, X_p)$ denote a random 1 x p vector of explanatory variables or features, and let $Y \in \{1, \ldots, K\}$ denote the unknown component or cluster label. Given a sample of X values, the goal is to estimate the number of clusters K and to estimate, for each observation, its cluster label Y.

Suppose we have data $\mathbf{X} = (x_{ij})$ on p explanatory variables (for example, genes) for n observations (for example, tumor mRNA samples), where x_{ij} denotes the realization of variable X_j for observation i and $\mathbf{x}_i = (x_{i1},\ldots,x_{ip})$ denotes the data vector for observation i, $i = 1,\ldots,n$, $j = 1,\ldots,p$. We consider clustering procedures that partition the learning set $\mathcal{L} = \{\mathbf{x}_1,\ldots, \mathbf{x}_n\}$ into K clusters of observations that are 'similar' to each other, where K is a user-prespecified integer. More specifically, the clustering $\mathcal{P}(\cdot;\mathcal{L})$ assigns class labels $\mathcal{P}(\mathbf{x}_i;\mathcal{L}) = \hat{y}_i$ to each observation, where $\hat{y}_i \in \{1,\ldots,K\}$. Clustering procedures generally operate on a matrix of pairwise dissimilarities (or similarities) between the observations to be clustered, such as the Euclidean or Manhattan distance matrices [8]. A partitioning of the learning set can be produced directly by partitioning clustering methods (for example, k-means, partitioning around medoid (PAM), self-organizing maps (SOM)) or by hierarchical clustering methods, by 'cutting' the dendrogram to obtain K 'branches' or clusters. Important issues, which will only be addressed briefly in this article, include: the selection of observational units, the selection of variables for defining the groupings, the transformation and standardization of variables, the choice of a similarity or dissimilarity measure, and the choice of a clustering method [9]. Our main concern here is to estimate the number of clusters K.

When a clustering algorithm is applied to a set of observations, a partition of the data is returned whether or not the data show a true clustering structure, that is, whether or not $K = 1$. This fact causes no problems if clustering is done to obtain a practical grouping of the given set of objects, as for organizational or visualization purposes (for example, hierarchical clustering for displaying large gene-expression data matrices as in Eisen *et al.* [10]). However, if interest lies primarily in the recognition of an unknown classification of the data, an artificial clustering is not satisfactory, and clusters resulting from the algorithm must be investigated for their relevance and reproducibility. This task can be carried out by descriptive and graphical exploratory methods, or by relying on probabilistic models and suitable statistical significance tests (for example [11,12]).

We argue here that validating the results of a clustering procedure can be done effectively by focusing on prediction accuracy. Once new classes are identified and class labels are assigned to the observations, the next step is often to build a classifier for predicting the class of future observations. The reproducibility or predictability of cluster assignments becomes very important in this context, and therefore provides a motivation for using ideas from supervised learning in an unsupervised setting. Resampling methods such as bagging [13] and boosting [14,15] have been applied successfully in the field of supervised learning to improve prediction accuracy. We propose here a novel resampling method, Clest, which combines ideas from discriminant and cluster analysis for estimating the number of clusters in a dataset. Although the proposed resampling methods are applicable to general clustering problems and procedures, particular attention is given to the clustering of tumors on the basis of gene-expression data using the partitioning around medoids (PAM) procedure (see below).

Partitioning around medoids

The new Clest procedure is demonstrated using the PAM method of Kaufman and Rousseeuw [16]. As implemented in the cluster package in R and S-Plus, the PAM function takes as its arguments a dissimilarity matrix (for example the Euclidean distance matrix as used here) and a prespecified

number of clusters K. The PAM procedure is based on the search for K representative objects, or medoids, among the observations to be clustered. After finding a set of K medoids, K clusters are constructed by assigning each observation to the nearest medoid. The goal is to find K medoids that minimize the sum of the dissimilarities of the observations to their closest medoid. The algorithm first looks for a good initial set of medoids, then finds a local minimum for the objective function, that is, a solution such that there is no single switch of an observation with a medoid that will decrease the objective.

The PAM method tends to be more robust and computationally efficient than k-means. In addition, PAM provides a graphical display, the silhouette plot, which can be used to select the number of clusters and to assess how well individual observations are clustered. Let a_i denote the average dissimilarity between i and all other observations in the cluster to which i belongs. For any other cluster C, let $d(i,C)$ denote the average dissimilarity of i to all objects of C and let b_i denote the smallest of these $d(i,C)$. The silhouette width of observation i is $sil_i = (b_i - a_i)/\max(a_i,b_i)$ and the overall average silhouette width is simply the average of sil_i over all observations i, $\bar{sil} = \sum_i sil_i/n$. Intuitively, objects with large silhouette width sil_i are well clustered, whereas those with small sil_i tend to lie between clusters. Kaufman and Rousseeuw suggest estimating the number of clusters K by that which gives the largest average silhouette width, $\bar{sil}$.

Existing methods for estimating the number of clusters in a dataset

Null hypothesis

Suppose that the maximum possible number of clusters in the data is set to M, $2 \le M \le n$. One approach to estimating the number of clusters K is to look for $\hat{K}$, $1 < \hat{K} \le M$, that provides the strongest significant evidence against the null hypothesis H_0 of $K = 1$, that is, 'no clusters' in the data. Two commonly used parametric null hypotheses are the unimodality hypothesis and the uniformity hypothesis.

Under the unimodality hypothesis, the data are thought to be a random sample from a multivariate normal distribution. This model typically gives a high probability of rejection of the null $K = 1$ if the data are sampled from a distribution with a lower kurtosis than the normal distribution, such as the uniform distribution [17].

The uniformity hypothesis, also referred to as random position hypothesis, states that the data are sampled from a uniform distribution in p-dimensional space [18-20]. Methods based on the uniformity hypothesis tend to be conservative, that is, lead to few rejections of the null hypothesis, when the data are sampled from a strongly unimodal distribution such as the normal distribution. In two or more dimensions, and depending on the test statistic, the results can be very sensitive to the region of support of the reference distribution [17].

For both types of hypotheses, evidence against the null hypothesis can be summarized formally under probability models for the data or more informally by using internal indices as described next.

Internal indices

Numerous methods have been proposed for testing the null hypothesis $K = 1$ and estimating the number of clusters in a dataset, however, none of them is completely satisfactory. Jain and Dubes [20] provide a general overview of such methods. The majority of existing approaches do not attempt to formally test the null hypothesis that $K = 1$, but rather look for the clustering structure under which a summary statistic of interest is optimal, being large or small depending on the statistic [21-23]. These statistics are typically functions of the within-clusters, and possibly between-clusters, sums of squares. They are referred to as internal indices, in the sense that they are computed from the same observations that are used to create the clustering. Consequently, the distribution of these indices is intractable. In particular, as clustering methods attempt to maximize the separation between clusters, the ordinary significance tests such as analysis of variance F-tests are not valid for testing differences between the clusters. Milligan and Cooper [12] conducted an extensive Monte Carlo evaluation of 30 internal indices. Other approaches include modeling the data using Gaussian mixtures and applying a Bayesian criterion to determine the number of components in the mixture [11]. A recent proposal of Tibshirani *et al.* [24], called the gap statistic method, calibrates an internal index, such as the within-clusters sum of squares, against its expectation under a suitably defined null hypothesis (note that gap tests have been used in another context in cluster analysis by Bock [18] to test the null hypothesis of a 'homogeneous' population against the alternative of 'heterogeneity'). Tibshirani *et al.* carried out a comparative Monte Carlo study of the gap statistic and several of the internal indices that showed a better performance in the study of Milligan and Cooper [12]. These internal indices and the gap statistic are described in more detail below.

For a given partition of the learning set into $1 \le k \le M$ clusters, define $\mathbf{B}_k$ and $\mathbf{W}_k$ to be the p x p matrices of between and within k-clusters sums of squares and cross-products [8]. Note that $\mathbf{B}_1$ is not defined. The following six internal indices are commonly used to estimate the number of clusters in a dataset**.**

sil: Kaufman and Rousseeuw [16] suggest selecting the number of clusters $k \ge 2$ which gives the largest average silhouette width, $\bar{sil}_k$. Silhouette widths were defined above with the clustering procedure PAM.

ch: Calinski and Harabasz [21]. For each number of clusters $k \geq 2$, define the index

$$ch_k = \frac{\mathrm{tr}\mathbf{B}_k/(k-1)}{\mathrm{tr}\mathbf{W}_k/(n-k)},$$

where tr denotes the trace of a matrix, that is, the sum of the diagonal entries. The estimated number of clusters is $\mathrm{argmax}_{k \geq 2}\, ch_k$.

kl: Krzanowski and Lai [23]. For each number of clusters $k \geq 2$, define the indices

$$diff_k = (k-1)^{2/p}\,\mathrm{tr}\mathbf{W}_{k-1} - k^{2/p}\,\mathrm{tr}\mathbf{W}_k \quad \text{and}$$

$$kl_k = |\, diff_k\,|/|diff_{k+1}|.$$

The estimated number of clusters is $\mathrm{argmax}_{k \geq 2}\, kl_k$.

hart: Hartigan [25]. For each number of clusters $k \geq 1$, define the index

$$hart_k = \left(\frac{\mathrm{tr}\mathbf{W}_k}{\mathrm{tr}\mathbf{W}_{k+1}} - 1\right)(n-k-1).$$

The estimated number of clusters is the smallest $k \geq 1$ such that $hart_k \leq 10$.

gap or **gapPC**: Tibshirani *et al.* [24]. This method compares an observed internal index, such as the within-clusters sum of squares, to its expectation under a reference null distribution as follows. For each number of clusters $k \geq 1$, compute the within-clusters sum of squares $\mathrm{tr}\mathbf{W}_k$. Generate B (here $B = 10$) reference datasets under the null distribution and apply the clustering algorithm to each, calculating the within-clusters sums of squares $\mathrm{tr}\mathbf{W}_k{}^1, \ldots, \mathrm{tr}\mathbf{W}_k{}^B$. Compute the estimated gap statistic

$$gap_k = \frac{1}{B}\sum_b \log \mathrm{tr}\,\mathbf{W}_k^b - \log \mathrm{tr}\,\mathbf{W}_k$$

and the standard deviation sd_k of $\log \mathrm{tr}\mathbf{W}_k{}^b$, $1 \leq b \leq B$. Let $\tilde{sd}_k = sd_k\sqrt{[(1 + 1/B)]}$. The estimated number of clusters is the smallest $k \geq 1$ such that $gap_k \geq gap_k^* - \tilde{sd}_k^*$, where $k^* = \mathrm{argmax}_{k \geq 1}\, gap_k$.

Tibshirani *et al.* [24] chose the uniformity hypothesis to create a reference null distribution and considered two approaches for constructing the region of support of the distribution. In the first approach, the sampling window for the jth variable, $1 \leq j \leq p$, is the range of the observed values for that variable. In the second approach, following Sarle [17], the variables are sampled from a uniform distribution over a box aligned with the principal components of the centered design matrix (that is, the columns of $\mathbf{X}$ are first set to have mean 0 and the singular value decomposition of $\mathbf{X}$ is computed). The new design matrix is then back-transformed to obtain a reference dataset. Whereas the first approach has the advantage of simplicity, the second takes into account the shape of the data distribution. Note that in both approaches the variables are sampled independently. The version of the gap method that uses the original variables to construct the region of support is referred to as gap and the second version as gapPC, where 'PC' stands for principal components.

Note that of the above methods, only hart, gap, and gapPC allow the estimation of only one cluster in the data, that is, $\hat{K} = 1$.

External indices

The term 'validation of a clustering procedure' usually refers to the ability of a given method to recover the true clustering structure in a dataset. There have been several attempts to assess validity on theoretical grounds [18,25]; however, such approaches turn out to be of little applicability in the context of high-dimensional complex datasets. In many validation studies, clustering methods are evaluated on their performance on empirical datasets with *a priori* known cluster labels [25] or, more commonly, on simulation studies where true cluster labels are known. To assess the ability of a clustering procedure to recover true cluster labels it is necessary to define a measure of agreement between two partitions; the first partition being the *a priori* known clustering structure of the data, and the second partition resulting from the clustering procedure. In the clustering literature, measures of agreement between partitions are referred to as external indices; several such indices are reviewed next.

Consider two partitions of n objects $\mathbf{x}_1, \ldots, \mathbf{x}_n$: the R-class partition $\mathcal{U} = \{u_1, \ldots, u_R\}$ and the C-class partition $\mathcal{V} = \{v_1, \ldots, v_C\}$. External indices of partitional agreement can be expressed in terms of a contingency table (Table 1), with entry n_{ij} denoting the number of objects that are both in clusters u_i and v_j, $i = 1,\ldots,R$, $j = 1,\ldots,C$ [20]. Let $n_{i.} = \sum_{j=1}^{C} n_{ij}$ and $n_{.j} = \sum_{i=1}^{R} n_{ij}$ denote the row and column sums of the contingency table, respectively, and let $Z = \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}^2$.

Table 1

Contingency table for two partitions of *n* objects

	v_1	v_2	$\cdots$	v_C	
u_1	n_{11}	n_{12}	$\cdots$	n_{1C}	$n_{1.}$
u_2	n_{21}	n_{22}	$\cdots$	n_{2C}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
u_R	n_{R1}	n_{R2}	$\cdots$	n_{RC}	$n_{R.}$
	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.C}$	$n_{..} = n$

The following indices can then be used.

1. **Rand:** Rand [26]

$$Rand = 1 + \left(Z - (1/2)\left(\sum_{i=1}^{R} n_{i.}^2 + \sum_{j=1}^{C} n_{.j}^2\right)\right)/\binom{n}{2} \simeq$$

2. **Jaccard**: Jain and Dubes [20]

$$Jac = \left(Z - n\right)/\left(\sum_{i=1}^{R} n_{i.}^2 + \sum_{j=1}^{C} n_{.j}^2 - Z - n\right).$$

3. **FM:** Fowlkes and Mallows [27]

$$FM = \left(1/2\right)\left(Z - n\right)/\left[\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}\right]^{1/2}.$$

Note that Rand and FM are linear functions of Z, and hence are linear functions of one another, conditional on the row and column sums in Table 1. If the row and column sums in Table 1 are fixed, but the partitions are selected at random; that is, if there is independence in the table, the hypergeometric distribution can be applied to determine the expected value of quantities such as Z. In particular

$$E\left[\sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2}\right] = (1/2)E(Z - n) = \sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}/\binom{n}{2}.$$

An external index S is often standardized in such a way that its expected value is 0 when the partitions are selected at random and 1 when they match perfectly. This amounts to computing a standardized external index

$$S' = \frac{S - E(S)}{S_{max} - E(S)},$$

where $S_{\max}$ is the maximum value of the statistic S and $E(S)$ is the expected value of S when partitions are selected at random. Accordingly, an often used correction for the Rand statistic is

$$Rand' = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2} - \left[1/\binom{n}{2}\right]\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}}{(1/2)\left[\sum_{i=1}^{R}\binom{n_{i.}}{2} + \sum_{j=1}^{C}\binom{n_{.j}}{2}\right] - \left[1/\binom{n}{2}\right]\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}}.$$

The significance of an observed external index is usually assessed under the assumption that the two partitions to be compared are independent. This assumption does not hold for the resampling methods described in the following section, since the same data are used to produce the two partitions. Nevertheless, external indices are convenient tools for comparing two clusterings, and are used in the new resampling method Clest. In this context, one should think of these indices as internal rather than external measures.

Results

Clest, a prediction-based resampling method for estimating the number of clusters

We propose a new prediction-based resampling method, Clest, for estimating the number of clusters, if any, in a dataset. The idea behind Clest is very intuitive if one is concerned with reproducibility or predictability of cluster assignments.

It is proposed to estimate the number of clusters K by repeatedly randomly dividing the original dataset into two non-overlapping sets, a learning set $\mathcal{L}^b$ and a test set $\mathcal{T}^b$. For each iteration and for each number of clusters k, a clustering $\mathcal{P}(\cdot;\mathcal{L}^b)$ of the learning set $\mathcal{L}^b$ is obtained and a predictor $C(\cdot;\mathcal{L}^b)$ is built using the class labels from the clustering. The predictor $C(\cdot;\mathcal{L}^b)$ is then applied to the test set $\mathcal{T}^b$ and the predicted labels are compared to those produced by applying the clustering procedure to the test set, using one of the external indices (or similarity statistics) described in the Background section. The number of clusters is estimated by comparing the observed similarity statistic for each k to its expected value under a suitable null distribution with $K = 1$. The estimated number of clusters is defined to be the $\hat{K}$ corresponding to the largest significant evidence against the null hypothesis of $K = 1$.

An early version of this approach was introduced by Breckenridge [28] under the name of replication analysis and was designed to evaluate the stability of a clustering. In the original replication analysis, the number of clusters k is fixed, and the data are randomly divided into two samples. A clustering procedure partitions both samples into k clusters, and the centroids of the clusters of the first sample are computed. A second set of labels is assigned to the observations in the second sample by assigning to each observation the cluster label of the closest centroid from the first sample. Finally, an external index is used to assess the agreement between the two partitions of the second sample. This measure reflects the stability of the clustering structure. The Clest procedure proposed here generalizes the work of Breckenridge [28].

Clest procedure for estimating the number of clusters in a dataset

Denote the maximum possible number of clusters by M, $2 \le M \le n$. For each number of clusters k, $2 \le k \le M$, perform steps 1-4.

1. Repeat the following B times:
   (a) Randomly split the original learning set $\mathcal{L}$ into two non-overlapping sets, a learning set $\mathcal{L}^b$ and a test set $\mathcal{T}^b$.
   (b) Apply a clustering procedure $\mathcal{P}$ to the learning set $\mathcal{L}^b$ to obtain a partition $\mathcal{P}(\cdot;\mathcal{L}^b)$.
   (c) Build a classifier $C(\cdot;\mathcal{L}^b)$ using the learning set $\mathcal{L}^b$ and its cluster labels.
   (d) Apply the resulting classifier to the test set $\mathcal{T}^b$.

(e) Apply the clustering procedure $\mathcal{P}$ to the test set $\mathcal{T}^b$ to obtain a partition $\mathcal{P}(\cdot;\mathcal{T}^b)$.

(f) Compute an external index $s_{k,b}$ comparing the two sets of labels for $\mathcal{T}^b$, namely the labels obtained by clustering and prediction.

2. Let t_k = median($s_{k,1}$, …, $s_{k,B}$) denote the observed similarity statistic for the k-cluster partition of the data.

3. Generate B_0 datasets under a suitable null hypothesis. For each reference dataset, repeat the procedure described in steps 1 and 2 above, to obtain B_0 similarity statistics $t_{k,1}$, …, $t_{k,B0}$.

4. Let t_k^0 denote the average of these B_0 statistics, $t_k^0 = [1/(B_0)]\sum_{b=1}^{B_0} t_{k,b}$, and let p_k denote the proportion of the $t_{k,b}$, $1 \le b \le B_0$, that are at least as large as the observed statistic t_k, that is, the p-value for t_k. Finally, let $d_k = t_k - t_k^0$ denote the difference between the observed similarity statistic and its estimated expected value under the null hypothesis of $K = 1$.

Define the set K as

$$K^- = \{2 \le k \le M : p_k \le p_{max}, d_k \ge d_{min}\},$$

where p_{max} and d_{min} are preset thresholds (see Parameters of the Clest procedure section below). If this set is empty, estimate the number of clusters as $\hat{K} = 1$. Otherwise, let $\hat{K} = \text{argmax}_{k \in K^-} d_k$, that is, take the number of clusters $\hat{K}$ that corresponds to the largest significant difference statistic d_k.

Parameters of the Clest procedure

In this paper, the following decisions are made regarding the different parameters for the Clest procedure (see summary in Table 2).

Clustering procedure: partitioning around medoids (PAM)

The PAM clustering procedure of Kaufman and Rousseeuw [16], implemented in the cluster package in R and S-Plus, was used to cluster observations based on the Euclidean distance metric (see Background).

Classifier: diagonal linear discriminant analysis (DLDA)

For multivariate Gaussian class conditional densities, that is, for $\mathbf{x}|y = k \sim N(\mu_k,\Sigma_k)$, the maximum likelihood (ML) discriminant rule (or Bayes rule with uniform class priors) predicts the class of an observation $\mathbf{x}$ by that which gives the largest likelihood to $\mathbf{x}$, that is,

$$C(x) = \text{argmin}_{1\le k\le K}\left\{(x - \mu_k)\Sigma_k^{-1}(x - \mu_k)' + \log|\Sigma_k|\right\}.$$

When the class densities have the same diagonal covariance matrix $\Sigma = \text{diag}(\sigma_1^2,\ldots,\sigma_p^2)$, the discriminant rule is linear and given by

Table 2

Parameters for Clest

Clest parameter	Value
Maximum number of clusters	M = 10 for microarray data
	M = 5 for simulated data
Number of learning/test set iterations	B = 20
Number of reference datasets	B_0 = 20
Size of learning sets $\mathcal{L}^b$	$2n/3$
Clustering procedure	PAM
Classifier	Linear discriminant analysis with diagonal covariance matrix - DLDA
Reference null distribution	Uniformity hypothesis
External index	Fowlkes and Mallows [27] external index, FM
Maximum p-value	p_{max} = 0.05
Minimum difference statistic	d_{min} = 0.05

$$C(x) = \text{argmin}_{1\le k\le K}\sum_{j=1}^{p}\frac{(x_j - \mu_{kj})^2}{\sigma_j^2}.$$

For the corresponding sample ML discriminant rules, the population mean vectors and covariance matrices are estimated from a learning set by the sample mean vectors and covariance matrices, respectively: $\hat{\mu}_k = \bar{\mathbf{x}}_k$ and $\hat{\mathbf{\Sigma}}_k = \mathbf{S}_k$. For the constant covariance matrix case, the pooled estimate of the common covariance matrix is used: $\hat{\mathbf{\Sigma}} = \sum_k(n_k - 1)\,\mathbf{S}_k/(n - K)$, where n_k denotes the number of observations in class k and n is the total sample size. DLDA is a very simple classifier but it has been shown to perform well in complex situations, in particular, in an extensive study of discrimination methods for the classification of tumors using gene-expression data [29]. DLDA is also known as naive Bayes classification.

Reference null distribution

The reference datasets are generated under the uniformity hypothesis as in the gap statistic method (see Background).

External index

All the external indices described in Background were considered. The FM index [27] was found to be superior to the other indices when reference datasets are generated under the uniformity hypothesis (data not shown).

Threshold parameters, p_{max} and d_{min}

The choices p_{max} = 0.05 and d_{min} = 0.05 are *ad hoc* and can probably be improved upon. Nevertheless, this rule gives a satisfactory performance and is used in the absence of a better choice.

Number of iterations and reference datasets

Here we used $B = B_0 = 20$. In general, the Clest procedure is robust to the choice of B and B_0 (data not shown).

Comparison of procedures on simulated data

The new procedure Clest was compared to six existing methods presented in Background using data simulated from the models described in Materials and methods. Figure 1 displays bar plots for the percentage of simulations for which a given method correctly recovered the number of clusters for each of the eight models. Table 3 provides a more detailed account of the simulation results for each procedure. It can be seen that Clest gave uniformly good results over the range of models, its worst performance being for Model 7 with two overlapping clusters. The rest of the methods failed for at least one of the eight models considered. The gap procedure failed twice (Models 5 and 6) and gapPC failed once (Model 6). Neither gap nor gapPC were able to identify the presence of the two clusters for Model 6, which is a model with two drawn-out clusters and seven noise variables with varying variances. Both gap and gapPC consistently estimated one cluster for this model, perhaps because both methods are based on the within-clusters sums of squares and consequently are more affected by the variables with larger variances. In a majority of the simulations from Model 7, Clest, gap, and gapPC failed to distinguish between one and two clusters, while the simple hart index performed well. The rest of the procedures do not have, by definition, the ability to estimate one cluster and hence generally identified the two clusters. Interestingly, for Model 8 with three overlapping clusters, sil and ch performed poorly, choosing two clusters in a majority of the simulations, while hart and Clest showed the best performance. Overall, most methods tended to underestimate more often than they overestimated the number of clusters, but the situation was reversed for hart and kl. For Model 1 it is only fair to

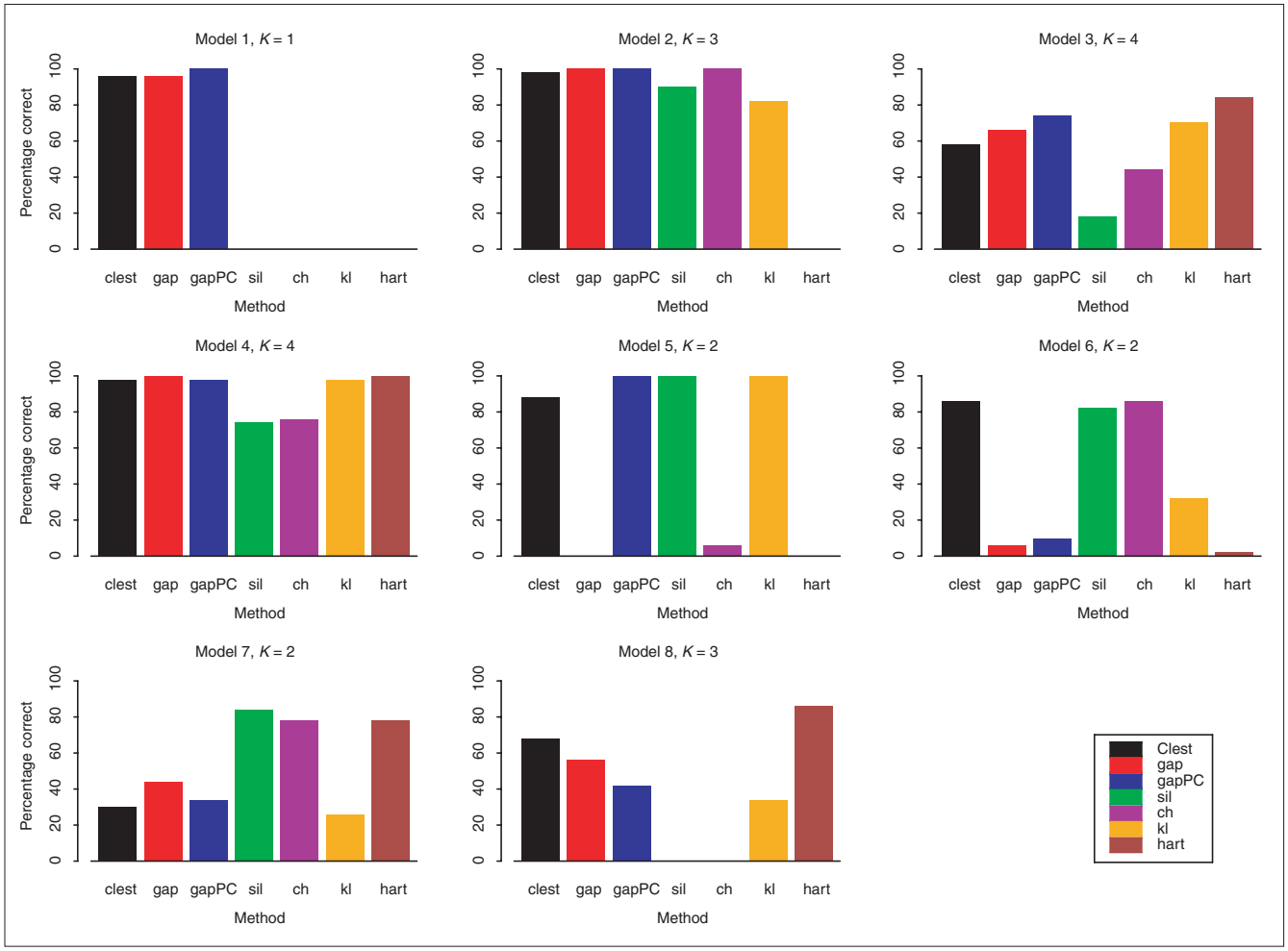

Figure 1

Estimating the number of clusters; results for simulated data. For each of the eight simulation models, the bar plots represent the percentage of simulations for which the number of clusters was correctly estimated by each method (out of 50 simulations).

Table 3

Estimating the number of clusters in simulated data

Method	Number of clusters, $\hat{K}$					
Model 1						
	1*	2	3	4	5	>5
Clest	**48**	2	0	0	0	0
gap	**48**	0	1	1	0	0
gapPC	**50**	0	0	0	0	0
sil	-	**37**	6	4	3	0
ch	-	**42**	7	1	0	0
kl	-	**12**	**14**	**11**	**13**	0
hart	0	5	**22**	16	7	0
Model 2						
	1	2	3*	4	5	>5
Clest	0	1	**49**	0	0	0
gap	0	0	**50**	0	0	0
gapPC	0	0	**50**	0	0	0
sil	-	5	**45**	0	0	0
ch	-	0	**50**	0	0	0
kl	-	0	**41**	2	7	0
hart	0	0	0	2	2	**46**
Model 3						
	1	2	3	4*	5	>5
Clest	0	1	20	**29**	0	0
gap	0	1	16	**33**	0	0
gapPC	0	1	12	**37**	0	0
sil	-	17	**24**	9	0	0
ch	-	8	**20**	**22**	0	0
kl	-	3	11	**35**	1	0
hart	0	0	8	**42**	0	0
Model 4						
	1	2	3	4*	5	>5
Clest	0	0	1	**49**	0	0
gap	0	0	0	**50**	0	0
gapPC	0	0	1	**49**	0	0
sil	-	5	8	**37**	0	0
ch	-	5	7	**38**	0	0
kl	-	0	1	**49**	0	0
hart	0	0	0	**50**	0	0
Model 5						
	1	2*	3	4	5	>5
Clest	0	**44**	0	6	0	0
gap	0	0	0	19	**31**	0
gapPC	0	**50**	0	0	0	0
sil	-	**50**	0	0	0	0
ch	-	3	0	**47**	0	0
kl	-	**50**	0	0	0	0
hart	0	0	0	0	0	**50**

Table 3 *(continued)*

Method	Number of clusters, $\hat{K}$					
Model 6						
	1	2*	3	4	5	>5
Clest	0	**43**	7	0	0	0
gap	**47**	3	0	0	0	0
gapPC	**43**	5	1	1	0	0
sil	-	**41**	5	4	0	0
ch	-	**43**	5	2	0	0
kl	-	**16**	9	**17**	8	0
hart	0	1	0	5	14	**30**
Model 7						
	1	2*	3	4	5	>5
Clest	**26**	15	6	3	0	0
gap	**25**	**22**	2	1	0	0
gapPC	**31**	17	2	0	0	0
sil	-	**42**	6	1	1	0
ch	-	**39**	10	0	1	0
kl	-	**13**	**15**	**10**	**12**	0
hart	6	**39**	5	0	0	0
Model 8						
	1	2*	3	4	5	>5
Clest	0	16	**34**	0	0	0
gap	0	22	**28**	0	0	0
gapPC	0	**28**	21	1	0	0
sil	-	**50**	0	0	0	0
ch	-	**50**	0	0	0	0
kl	-	**25**	17	4	4	0
hart	0	3	**43**	4	0	0

For each simulation model, the distribution of the estimated number of clusters is recorded for each method. The true number of clusters is denoted by the asterisk and the modes for the distribution of the 50 estimates are indicated in bold for each method. Note that sil, ch, and kl do not have the ability to estimate $\hat{K} = 1$ cluster.

compare Clest, gap, gapPC, and hart, as the other methods only estimate $\hat{K} \geq 2$.

In summary, for the simulation models considered here, Clest was the most robust and accurate, whereas hart performed worst. gapPC was better than gap and the rest of the methods performed similarly.

For a given model, it is of interest to consider the median value of the statistics used by each method to estimate the number of clusters. For each number of clusters k, the plots of the median values, over the 50 simulated datasets, of the Clest d_k-statistic, $gapPC_k$, and $\bar{sil}_k$ statistics are shown in Figures 2, 3 and 4, respectively. The d-statistic does not generally have local maxima except for Model 5. There, a local maximum appears at $K = 4$ clusters, but the global maximum occurs at $K = 2$. It can be seen that the ability of Clest to distinguish between one and two clusters is very low for Model 7; the median of the d_2 values is less than the significance cut-off d_{min} used in the Clest procedure. Indeed, the results in Table 3 show that Clest identified two clusters for only 30% of the datasets simulated from Model 7. The figures suggest that for the majority of the models, the global maximum of the median d_k-statistic is more pronounced than the global maxima of the median $gapPC_k$ and $\bar{sil}_k$ statistics, respectively. This again suggests good robustness and accuracy properties for the Clest method.

Comparison of procedures on microarray data

The new Clest method was also evaluated using gene-expression data from the four cancer microarray studies described in Materials and methods and summarized in Table 4. Recall that mRNA samples in the lymphoma, leukemia, and NCI60 datasets were assigned class labels from the laboratory

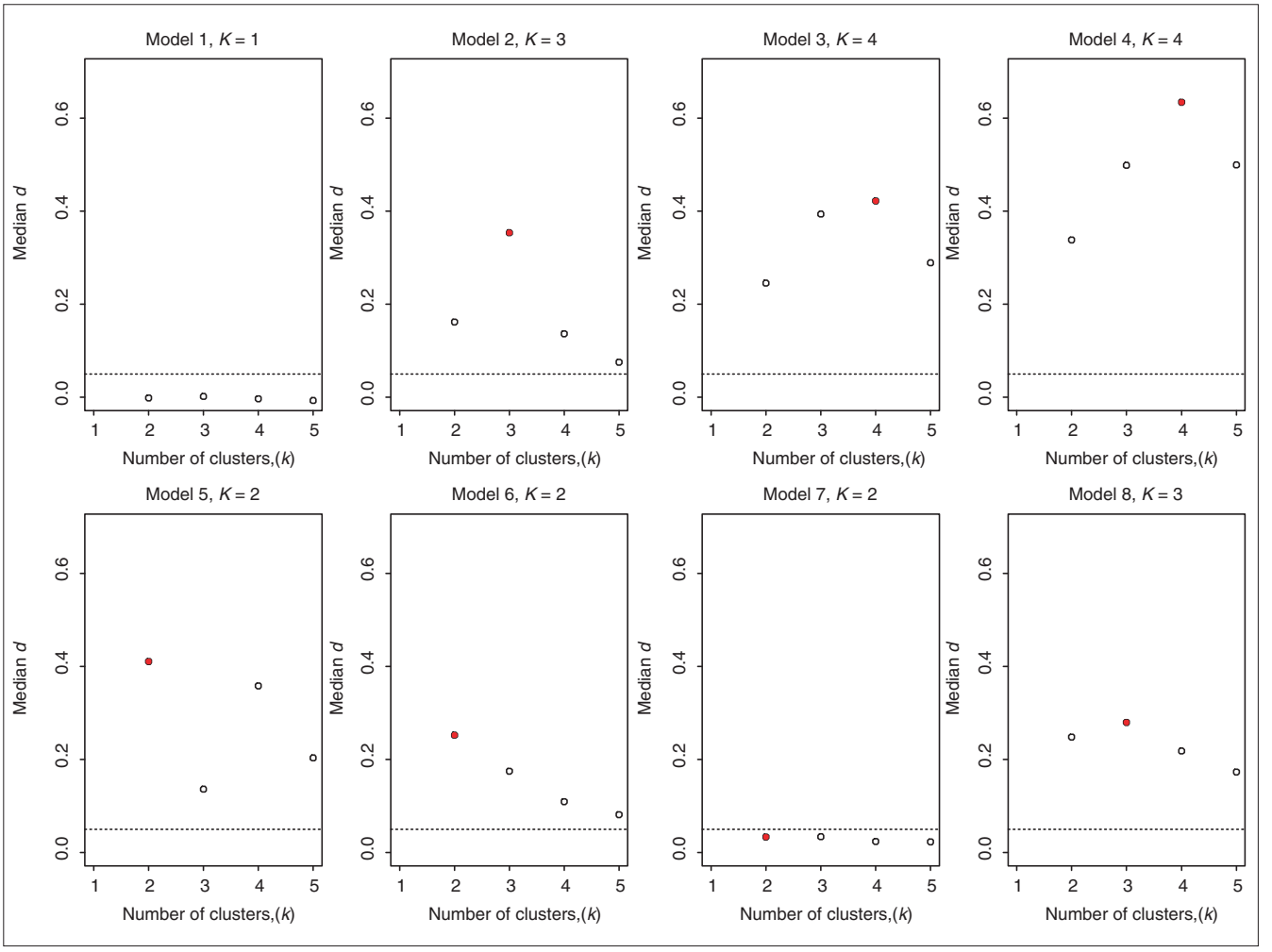

Figure 2
Estimating the number of clusters using the Clest procedure; results for simulated data. Plots are of median d_k versus k for each simulation model (medians are computed over 50 simulations). The horizontal line corresponds to the d_{min} cut-off of 0.05, and the true number of clusters is indicated by a filled plotting symbol.

analyses of the tumor samples or from *a priori* knowledge of the cell lines. For the melanoma dataset, tumor class labels were obtained from the statistical analysis described in Bittner *et al.* [30]. In the discussion that follows, these class labels are treated as known. The six methods described in Background and Clest were applied to estimate the number of clusters for each of the four microarray datasets; the results are presented in Table 5.

The methods Clest and sil correctly estimated the presumed number of classes for all but the NCI60 dataset, where both methods identified three clusters only. The gap and gapPC methods overestimated the number of clusters for all datasets, with the exception of gapPC, which identified eight clusters for the NCI60 dataset. The ch method estimated two clusters for each of the four datasets, whereas kl and hart identified four classes for the lymphoma dataset.

For Clest, gapPC, and sil, we further investigated how the strength of the evidence for the estimated number of clusters varied between datasets. Figure 5 displays plots of the d_k, $gapPC_k$, and $\bar{sil}_k$ statistics versus the number of clusters k. Error bars for d_k and $gapPC_k$ are based on the standard deviations of t_k and log tr$\mathbf{W}_k$ under their respective null distributions. Whereas the evidence for the existence of clusters is very strong for the lymphoma, leukemia, and NCI60 datasets, the evidence for the two clusters in the melanoma dataset is much weaker. In particular, for Clest, the maximum value of the d_k statistic barely reaches the d_{min} threshold of 0.05. For the leukemia dataset, the d_k statistic clearly peaks at $k = 3$ clusters and drops off abruptly; for the lymphoma and NCI60 datasets the decrease is more gradual. Note that according to Clest there was not enough evidence to identify the two DLBCL subclasses for the lymphoma dataset. Alizadeh *et al.* [1] identified these subclasses using subject matter

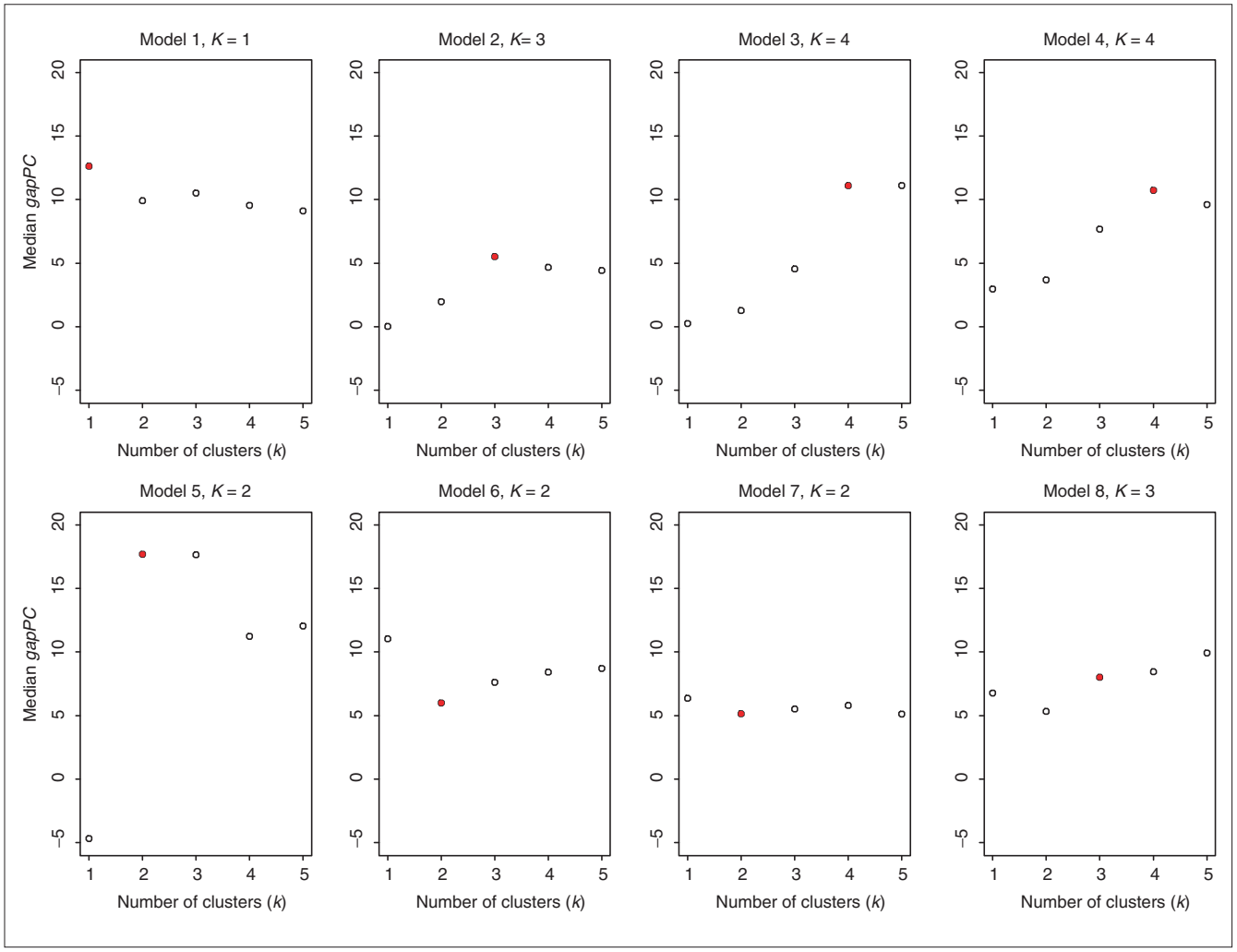

Figure 3
Estimating the number of clusters using the gapPC procedure; results for simulated data. Plots are of median $gapPC_k$ versus k for each simulation model (medians are computed over 50 simulations). The true number of clusters is indicated by a filled plotting symbol.

knowledge to select the genes for the clustering procedure; here the genes were selected in an unsupervised manner.

Discussion

Resampling methods such as bagging and boosting have been applied successfully in a supervised learning context to improve prediction accuracy. Here and in a related article [7], we have proposed resampling methods to address two main problems in cluster analysis: estimating the number of clusters, if any, in a dataset; improving and assessing the accuracy of a given clustering procedure. As the groups obtained from cluster analysis are often used later on for prediction purposes, the approaches to these two problems rely on and extend ideas from supervised learning. Although the methods are applicable to general clustering problems and procedures, particular attention is given to the clustering of tumors using gene-expression data. The performance of the proposed and existing methods was compared using simulated data and gene-expression data from four recently published cancer microarray studies.

To estimate the number of clusters in a dataset, we propose a prediction-based resampling method, Clest, which estimates the number of clusters K based on the reproducibility of cluster assignments. In comparative studies, Clest was generally found to be more accurate and robust than six existing methods. For the simulated datasets, Clest performed well across a wide range of models with varying numbers of overlapping and non-overlapping clusters, different numbers of variables and covariance matrix structures. Unlike methods based on between- or within-clusters sums of squares, the resampling method seems robust to the varying covariance structure of the variables.

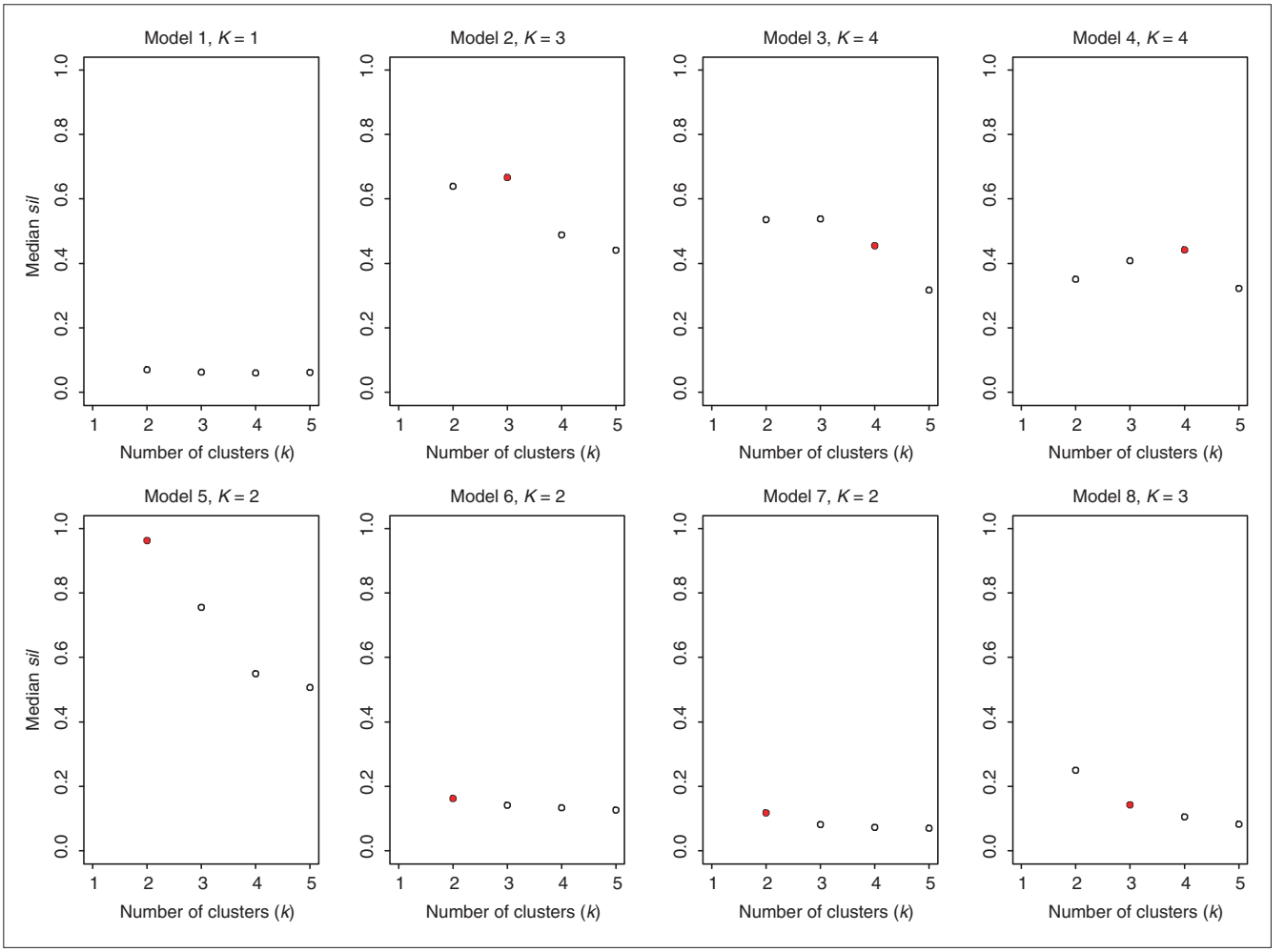

Figure 4
Estimating the number of clusters using the sil procedure; results for simulated data. Plots are of median $\bar{sil}_k$ versus k for each simulation model (medians are computed over 50 simulations). The true number of clusters is indicated by a filled plotting symbol.

For the microarray datasets, Clest and sil correctly estimated the number of tumor or cell-line clusters (as determined from *a priori* known or putative tumor and cell-line classes) for three out of the four datasets; the performance of other methods was significantly worse. We focus here on the clustering of tumor mRNA samples using gene-expression data. Once tumor classes are specified, an important next step would be the identification of marker genes that characterize these different tumor classes. A related question, which we have not considered here, is the 'transpose' clustering problem; that is, the clustering of genes that have similar expression levels across biological samples. One could then investigate the clusters for the presence of shared regulatory motifs among the genes [31]. This could lead to the identification of genes that are not only coexpressed but are also under similar regulatory control. Joint analysis of transcript level and sequence data should lead to greater biological insight into the molecular characterization of tumors.

A number of decisions were made regarding the different parameters of the Clest procedure. The clustering procedure PAM was used in the comparison; however, one should keep in mind that different clustering procedures can generate different partitions of the same data, possibly leading to different inferences about the number of clusters. In addition, the clustering (PAM) and prediction methods (DLDA or naive Bayes) considered in this article focus on similar features of the data, namely, the distance of the observations from cluster 'centers'. More work is needed to investigate the robustness of Clest to these choices. In particular, it would be interesting to consider prediction and clustering methods that focus on different aspects of the data (for example, classification trees instead of DLDA). Although it may seem that having a classifier as a parameter of the Clest procedure creates more room for error, we have found that this is not the case in practice. When the classifier in Clest performs poorly, other methods for estimating the number of clusters

Table 4

Description of microarray datasets

Dataset	Number of classes	Class sizes	Number of genes
Lymphoma* [1] (cDNA microarrays)	$K = 3$ classes	B-CLL (29) FL (9) DLBCL (43)	$p = 4{,}682$
Leukemia [3] (Affymetrix chips)	$K = 3$ classes	ALL B-cell (38) ALL T-cell (9) AML (25)	$p = 3{,}571$
NCI 60† [6] (cDNA microarrays)	$K = 8$ classes	Breast (7), CNS (6), colon (7), leukemia (6), melanoma (8), NSCLC (9), ovarian (6), renal (8)	$p = 5{,}244$
Melanoma‡ [30] (cDNA microarrays)	$K = 2$ classes	Group A (19) Group B (12)	$p = 3{,}613$

*The DLBCL class for the lymphoma dataset is likely to contain two subclasses.†For the NCI60 data, the two prostate cell lines and the unknown cell line (ADR-RES) were excluded from our analysis because of their small class size. ‡Note that for the first three datasets, tumor classes were known *a priori*, whereas for the melanoma dataset the two classes were inferred by Bittner *et al.* [30] by cluster analysis but not confirmed on an independent dataset.

Table 5

Estimating the number of clusters from microarray data

Dataset	Known	Clest	gap	gapPC	sil	ch	kl	hart
Lymphoma	**3**	3	10	8	3	2	4	4
Leukemia	**3**	3	10	5	3	2	3	3
NCI60	**8**	3	10	8	3	2	6	2
Melanoma	**2**	2	9	4	2	2	8	1

also perform poorly. Another important choice in the Clest procedure is the reference null distribution used to calibrate the observed similarity statistics t_k for different numbers of clusters. The uniformity hypothesis was used here; a natural alternative would be to consider random permutations of the variables, that is, permutations of the entries of the design matrix within columns. In Clest, the observed similarity statistics t_k are compared across numbers of clusters k by considering their distance from their estimated expected value t_k^0 under the null distribution. A more sensitive calibration may be achieved by taking scale into account, that is, by dividing the difference statistic d_k by the standard deviation of t_k under the null distribution, or even by considering p-values p_k for t_k. We briefly considered these refinements and found that on their own they did not allow good discrimination between the different ks. The Clest method does, however, use the idea of p-value in combination with the differences d_k, as it imposes an upper limit on the p-value p_k. Finally, the choice of cut-off parameters d_{min} and p_{max} was rather *ad hoc* and could be fine tuned.

We have not considered model-based methods, such as the Bayesian approach of Fraley and Raftery [11] or the mixture-model approach of McLachlan *et al.* [32]. Another issue only briefly addressed here is the selection of variables on which to base the clusterings. For the microarray datasets, genes were selected on the basis of the variance of their expression levels across samples, and it was found that the clusterings were fairly robust to the number of genes.

Resampling methods are promising tools for addressing various problems in cluster analysis. Ben-Hur *et al.* [33] have recently proposed a stability-based method for estimating the number of clusters, where stability is characterized by the distribution of pairwise similarities between clusterings obtained from subsamples of the data. It would be interesting to relate the approach of Ben-Hur *et al.* and Clest. Elsewhere, we proposed two bagged clustering methods for improving and assessing the accuracy of a given partitioning clustering procedure [7]. There, the bootstrap is used to generate and aggregate multiple clusterings and to assess the confidence of cluster assignments for individual observations. Leisch [34] proposed a bagged clustering method which is a combination of partitioning and hierarchical methods. A partitioning method is applied to bootstrap learning sets and the resulting partitions are combined by performing hierarchical clustering of the cluster centers. This method is similar in spirit to our two new bagging procedures [7].

Conclusions

Focusing on prediction accuracy in conjunction with resampling produces accurate and robust estimates of the number of clusters. As reproducibility of the cluster assignments is an integral part of the Clest method, the clustering results can be used reliably for building a classifier to predict the class of future observations. In addition, the procedure is robust to the covariance structure among variables.

Materials and methods

Simulation models

Procedures for estimating the number of clusters in a dataset were evaluated using simulated data from a variety of models, including those considered by Tibshirani *et al.* [24]. The models used for comparison contain different numbers of overlapping and non-overlapping clusters, different numbers of variables, and a wide range of covariance matrix structures. In addition, a variable number of irrelevant or 'noise' variables are included in the models. A noise variable is a variable whose distribution does not depend on the

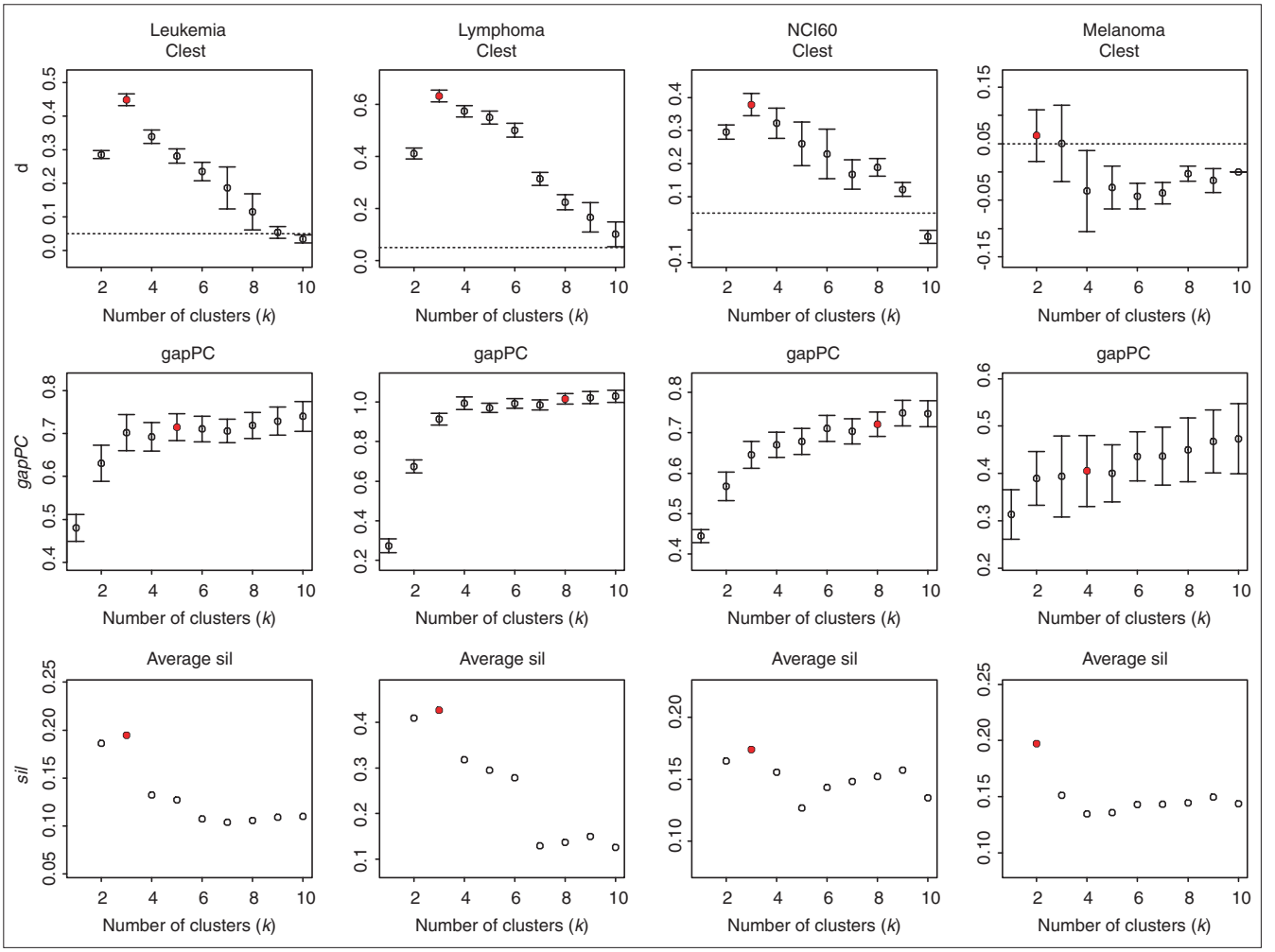

Figure 5
Estimating the number of clusters; results for microarray data. Plots of d_k, $gapPC_k$, and $\bar{sil}_k$ versus k, with error bars computed as described in Results. The horizontal lines for the d_k plots correspond to a d_{min} cut-off of 0.05. The estimates for the number of clusters are indicated by filled plotting symbols.

cluster label, and such variables are added to obscure the underlying clustering structure to be recovered.

Model 1. One cluster in 10 dimensions. n = 200 observations are simulated from the uniform distribution over the unit hypercube in p = 10 dimensions.

Model 2. Three clusters in two dimensions. The observations in each of the three clusters are independent bivariate normal random variables with means (0,0), (0,5), and (5,-3), respectively, and identity covariance matrix. There are 25, 25, and 50 observations in each of the 3 clusters, respectively.

Model 3. Four clusters in 10 dimensions, 7 noise variables. Each cluster is randomly chosen to have 25 or 50 observations, and the observations in a given cluster are independently drawn from a multivariate normal distribution with identity covariance matrix. For each cluster, the cluster means for the first three variables are randomly chosen from a $N(\mathbf{o}_3, 25\mathbf{I}_3)$ distribution, where $\mathbf{o}_p$ denotes a 1 x p vector of zeros and $\mathbf{I}_p$ denotes the p x p identity matrix. The means for the remaining seven variables are 0. Any simulation where the Euclidean distance between the two closest observations belonging to different clusters is less than 1 is discarded.

Model 4. Four clusters in 10 dimensions. Each cluster is randomly chosen to contain 25 or 50 observations, with means randomly chosen as $N(\mathbf{o}_{10}, 3.6\mathbf{I}_{10})$. The observations in a given cluster are independently drawn from a normal distribution with identity covariance matrix and appropriate mean vector. Any simulation where the Euclidean distance between the two closest observations belonging to different clusters is less than 1 is discarded.

Model 5. Two elongated clusters in three dimensions. Cluster 1 contains 100 observations generated as follows. Set

$x_1 = x_2 = x_3 = t$, with t taking on equally spaced values from -0.5 to 0.5. Gaussian noise with standard deviation of 0.1 is then added to each variable. Cluster 2 is generated in the same way except that the value 10 is added to each variable. This results in two elongated clusters, stretching out along the main diagonal of a three-dimensional cube, with 100 observations each.

Model 6. Two elongated clusters in 10 dimensions, 7 noise variables. The clusters are generated as in Model 5, but, in addition, seven noise variables are simulated independently from a normal distribution with mean 0 and variance v^2 for the vth variable, $4 \le v \le 10$.

Model 7. Two overlapping clusters in 10 dimensions, 9 noise variables. Each cluster contains 50 observations. The first variable in each of the two clusters is normally distributed with mean 0 and 2.5, respectively, and with variance 1. The remaining nine variables are simulated from the $N(\mathbf{0}_9,\mathbf{I}_9)$ distribution (independently of the first variable).

Model 8. Three overlapping clusters in 13 dimensions, 10 noise variables. Each cluster contains 50 observations. The first three variables have a multivariate normal distribution with mean vectors (0,0,0), (2,-2,2), and (-2,2,-2), respectively, and covariance matrix Σ, where $\sigma_{ii} = 1$, $1 \le i \le 3$, and $\sigma_{ij} = 0.5$, $1 \le i \ne j \le 3$. The remaining 10 variables are simulated independently from the $N(\mathbf{0}_{10},\mathbf{I}_{10})$ distribution.

Note that Models 1, 2, 4, and 5 were considered in Tibshirani *et al.* [24]. Model 3 is similar to the third model in [24], with the addition of seven noise variables. Model 6 is the same as Model 5, with the addition of seven noise variables.

Fifty datasets were simulated from each model and the methods described in the Background and Results sections were applied to estimate the number of clusters in the resulting datasets. We are primarily interested in comparing the percentage of simulations for which each procedure recovers the correct number of clusters, as this quantity reflects the accuracy of the procedure. However, for the purpose of future applications, it is useful to also know whether a method tends to underestimate or overestimate the true number of clusters. Hence, the full distribution of the number of clusters estimated by each method is presented in Table 3. Note that only the methods Clest, gap, gapPC and hart have the capability to identify one cluster in the data.

Microarray data

The new Clest procedure and existing methods described in Background were applied to gene-expression data from four recently published cancer microarray studies: the lymphoma dataset of Alizadeh *et al.* [1], the leukemia (ALL/AML) dataset of Golub *et al.* [3], the 60 cancer cell line (NCI60) dataset of Ross *et al.* [6], and the melanoma dataset of Bittner *et al.* [30] (see summary in Table 4). Note that the expression levels are, in general, highly processed data: the raw data in a microarray experiment consist of image files, and important pre-processing steps include image analysis of the scanned images and normalization. Because we chose to use publicly available datasets, most of these decisions were beyond our control, and one should bear in mind that different pre-processing decisions could have a large impact on the measured expression levels [35,36].

Lymphoma

This dataset comes from a study of gene expression in the three most prevalent adult lymphoid malignancies: B-cell chronic lymphocytic leukemia (B-CLL), follicular lymphoma (FL), and diffuse large B-cell lymphoma (DLBCL) (see [1,37] for a detailed description of the experiments). Gene-expression levels were measured using a specialized cDNA microarray, the Lymphochip, containing genes that are preferentially expressed in lymphoid cells or which are of known immunological or oncological importance. In each hybridization, fluorescent cDNA targets were prepared from a tumor mRNA sample (red-fluorescent dye, Cy5) and a reference mRNA sample derived from a pool of nine different lymphoma cell lines (green-fluorescent dye, Cy3). The cell lines in the common reference pool were chosen to represent diverse expression patterns, so that most spots on the array would exhibit a non-zero signal in the Cy3 channel. This study produced gene-expression data for $p = 4{,}682$ genes in $n = 81$ mRNA samples. The tumor mRNA samples consist of 29 cases of B-CLL, 9 cases of FL, and 43 cases of DLBCL. Alizadeh *et al.* [1] further showed that the DLBCL class is heterogeneous and comprises two distinct subclasses of tumors with different clinical behaviors. The gene-expression data are summarized by an 81 x 4,682 matrix $\mathbf{X} = (x_{ij})$, where x_{ij} denotes the base-2 logarithm of the Cy5/Cy3 background-corrected and normalized fluorescence intensity ratio for gene j in lymphoma sample i. The mean percentage of missing observations per array is 6.6% and missing data were inferred as outlined below. The data were standardized as described below.

Leukemia

The leukemia dataset is described in [3] and available at [38]**.** This dataset comes from a study of gene expression in two types of acute leukemia: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). Gene-expression levels were measured using Affymetrix high-density oligonucleotide arrays containing $p = 6{,}817$ human genes. The data comprise 47 cases of ALL (38 ALL B-cell and 9 ALL T-cell) and 25 cases of AML. Following Golub *et al.* (P. Tamayo, personal communication), three pre-processing steps were applied to the normalized matrix of intensity values available on the website (after pooling the 38 mRNA samples from the learning set and the 34 mRNA samples from the test set). First, a floor of 100 and ceiling of 16,000 was set; second, the data were filtered to exclude genes with max/min ≤ 5 or (max - min) ≤ 500, where max and min refer

respectively to the maximum and minimum intensities for a particular gene across the 72 mRNA samples; and third, the data were transformed to base 10 logarithms. The data are then summarized by a 72 x 3,571 matrix $\mathbf{X} = (x_{ij})$, where x_{ij} denotes the expression level for gene j in mRNA sample i. There are no missing values and the data were standardized as described below. Note that this standardization differs from the one described in Golub *et al.* [3].

NCI60

In this study, cDNA microarrays were used to examine the variation in gene expression among the 60 cell lines from the National Cancer Institute's (NCI60) anti-cancer drug screen [6,39]. The cell lines were derived from tumors with different sites of origin: 7 breast, 6 central nervous system (CNS), 7 colon, 6 leukemia, 8 melanoma, 9 non-small-cell-lung-carcinoma (NSCLC), 6 ovarian, 2 prostate, 8 renal, and 1 unknown (ADR-RES). Gene expression was studied using microarrays with 9,703 spotted DNA sequences. In each hybridization, fluorescent cDNA targets were prepared from a cell-line mRNA sample (red-fluorescent dye, Cy5) and a reference mRNA sample obtained by pooling equal mixtures of mRNA from 12 of the cell lines (green-fluorescent dye, Cy3). To investigate the reproducibility of the entire experimental procedure (cell culture, mRNA isolation, labeling, hybridization, scanning, and so on), a leukemia (K562) and a breast cancer (MCF7) cell line were analyzed by three independent microarray experiments. Ross *et al.* screened out genes with missing data in more than two arrays. In addition, because of their small class size, the two prostate cell lines and the unknown cell line (ADR-RES) were excluded from our analysis. The data are summarized by a 61 x 5,244 matrix $\mathbf{X} = (x_{ij})$, where x_{ij} denotes the base-2 logarithm of the Cy5/Cy3 background-corrected and normalized fluorescence intensity ratio for gene j in cell line i. The mean percentage of missing observations per array is 3.3% and missing data were inferred as outlined below. The data were standardized as described below.

Melanoma

The melanoma dataset is described in the recent paper of Bittner *et al.* [30] and is available at [40]. There are 31 melanoma samples and 7 control samples. Gene-expression levels were measured using cDNA microarrays with 8,150 probe sequences, representing 6,971 unique genes. In each hybridization, fluorescent cDNA targets were prepared from a melanoma or control mRNA sample (red-fluorescent dye, Cy5) and a common reference mRNA sample (green-fluorescent dye, Cy3). The following pre-processing steps were applied by Bittner *et al.* First, a gene was excluded from the analysis if its average mean intensity above background for the least intense signal (Cy3 or Cy5) across all experiments was $\leq$ 2,000 or its average spot size across all experiments was $\leq$ 30 pixels; and second, a floor and ceiling of 0.02 and 50, respectively, were applied to the individual intensity log-ratios. This initial screening resulted in a dataset of 3,613 genes (see Supplemental Information to [30], document II, page 2). Finally, Bittner *et al.* did not include the seven control samples in their analysis. The data are summarized by a 31 x 3,613 matrix $\mathbf{X} = (x_{ij})$, where x_{ij} denotes the base-2 logarithm of the Cy5/Cy3 background-corrected and normalized fluorescence intensity ratio for gene j in mRNA sample i. There were no *a priori* known classes for this dataset, but the analysis of Bittner *et al.* suggests that two classes may be present in the data, with observations in one of the classes (Group A in their figures) being more tightly clustered. There were no missing values and the data were standardized as described below. Note that this standardization is slightly different from the one described in [30].

Imputation of missing data

For the lymphoma and NCI60 datasets, each array contains a number of genes with fluorescence-intensity measurements that were flagged by the experimenter and recorded as missing data points. Missing data were imputed by a simple k-nearest-neighbor algorithm, in which the neighbors are the genes and the distance between neighbors is based on the correlation between their gene-expression levels across arrays. For each gene with missing data: first compute its correlation with all other p - 1 genes, and then, for each missing array, identify the k nearest genes having data for this array and infer the missing entry from the average of the corresponding entries for the k neighbors. A value of k = 5 neighbors was used for the lymphoma and NCI60 datasets. For a detailed study of methods for imputing missing values in microarray experiments, see [41], which suggests that a nearest-neighbor approach provides accurate and robust estimates of missing values.

Standardization

The gene-expression data were standardized so that the observations (arrays) have mean 0 and variance 1 across variables (genes). Standardizing the data in this fashion achieves a location and scale normalization of the different arrays. In a study of normalization methods, we have found scale adjustment to be desirable in some cases, to prevent the expression levels in one particular array from dominating the average expression levels across arrays [36]. Furthermore, this standardization is consistent with the common practice in microarray experiments of using the correlation between the gene-expression profiles of two mRNA samples to measure their similarity [1,4,6]. In practice, however, we recommend general adaptive and robust normalization methods which correct for intensity, spatial, and other types of dye biases using robust local regression [36].

Preliminary gene selection

Expression levels were monitored for thousands of genes in each of the four studies. However, the majority of the genes exhibit near-constant expression levels, as measured by the variance (or coefficient of variation) of the expression levels across tumor samples. Genes showing nearly constant

expression levels are not likely to be useful for classification purposes; therefore, we chose to exclude low-variance genes from the clustering process.

Figure 6 displays for each dataset the individual gene variances divided by the maximum variance over all genes. All four variance curves show a sharp drop-off which gradually flattens. The plots are remarkably similar for all the datasets, with the melanoma dataset having the fastest drop-off. In this report, the p = 100 most variable genes were used to analyze the leukemia, lymphoma and melanoma datasets, and the p = 200 most variable genes were used for the NCI60 dataset as it contains more classes. Increasing the number of genes to p = 300-400 or decreasing the number of genes to p = 50 did not have much effect on the results (data not shown). One could also select genes based on a coefficient of variation filter.

Correlation matrices

The following is not part of the cluster analysis *per se*, but is an interesting side-step which may be predictive of the results of the forthcoming analysis. Recall that for the first three datasets, tumor classes were known *a priori*, and for the melanoma dataset two classes were inferred by Bittner *et al*. [30] through cluster analysis. For each dataset, images of the n x n correlation matrix for the n mRNA samples are displayed in Figures 7-10, with observations grouped according to their *a priori* known or putative classes. Note that if observations are highly correlated within classes, the correlation image in this representation should show bright red squares along the diagonal.

Lymphoma

The existence of three well-separated classes for the lymphoma dataset is reflected in Figure 7 for both sets of genes, the classes being more clearly separated when the majority of the genes are screened out. Recall that gene-expression levels were measured using a specialized cDNA microarray, the Lymphochip, enriched in genes that are involved in the immune system. This may partly account for the clear separation of the classes even when the correlation matrix is computed using the full set of genes. When PAM is applied to the lymphoma dataset using the 100 genes with the largest variance, the K = 2, 3, 4, 5 partitions are as follows. For K = 2 classes, one cluster consists of the FL and DLBCL classes combined and the other consists of the CLL class. This could reflect differences in tissue sampling, as the CLL mRNA samples were obtained from peripheral blood cells as opposed to lymph-node biopsy specimens for the FL and DLBCL samples. For K = 3, all three classes (CLL, FL,

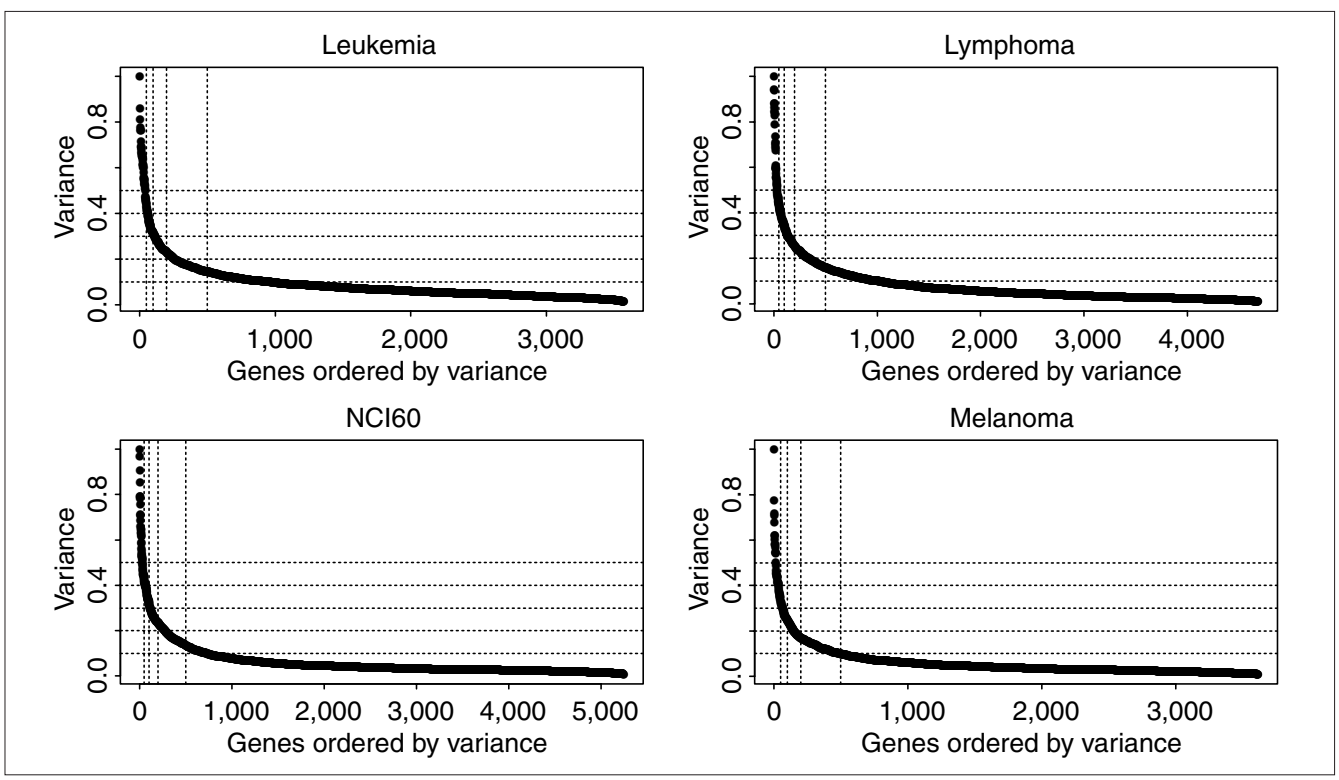

Figure 6
Gene variances for microarray datasets. Plots of the variance of the expression levels of each gene across mRNA samples. The variances are scaled by the maximum variance over all genes and the genes are ordered by variance in descending order. The vertical lines correspond to 50, 100, 200 and 500 genes, and the horizontal lines correspond to ratios of variances of 0.1, 0.2, 0.3, 0.4 and 0.5.

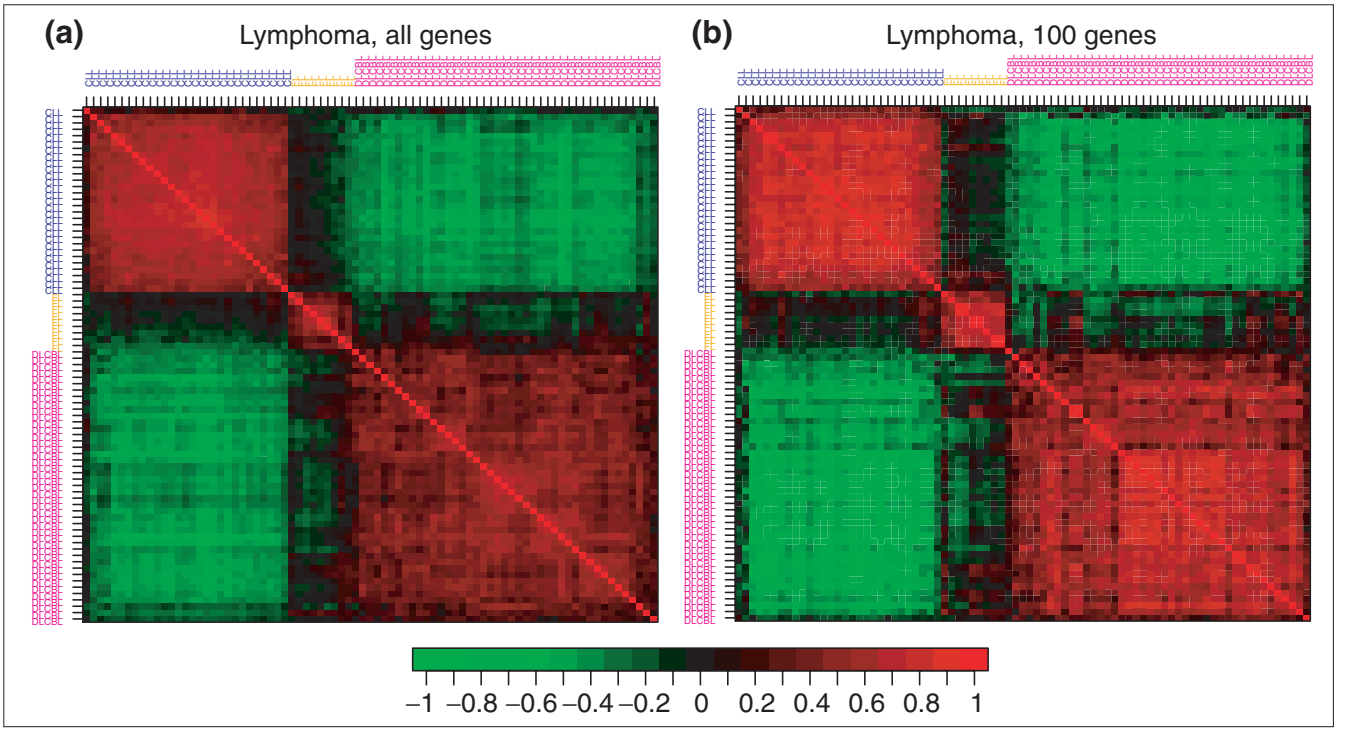

Figure 7
Correlation matrix, lymphoma dataset. Images of the correlation matrix for the 81 B-CLL, FL, and DLBCL samples based on expression profiles for **(a)** all p = 4,682 genes and **(b)** the p = 100 genes with the largest variance. The mRNA samples are ordered by class, first B-CLL (blue), then FL (orange), and finally DLBCL (magenta). Correlations of zero are represented in black, increasingly positive correlations are represented with reds of increasing intensity, and increasingly negative correlations are represented with greens of increasing intensity. The color bar below the images can be used for calibration purposes.

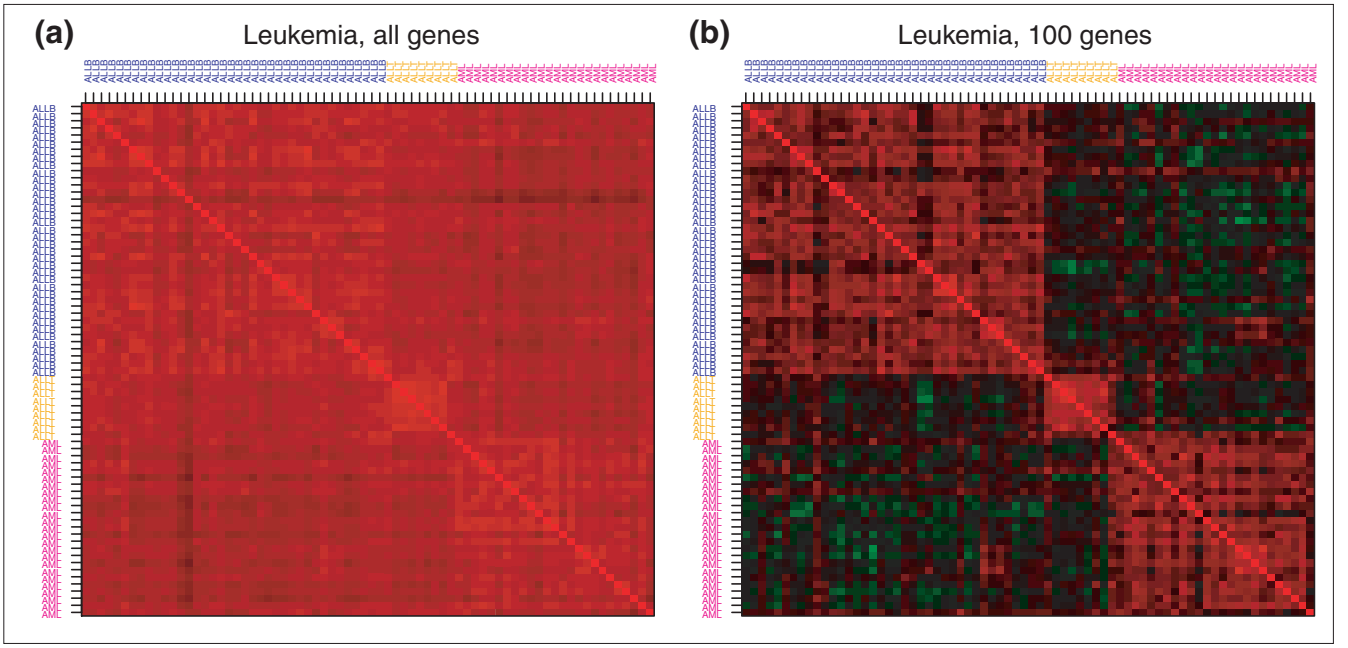

Figure 8
Correlation matrix, leukemia dataset. Images of the correlation matrix for the 72 ALL B-cell, ALL T-cell, and AML samples based on expression profiles for **(a)** all p = 3,571 genes and **(b)** the p = 100 genes with the largest variance. The mRNA samples are ordered by class, first ALL B-cell (blue), then ALL T-cell (orange), and finally AML (magenta).

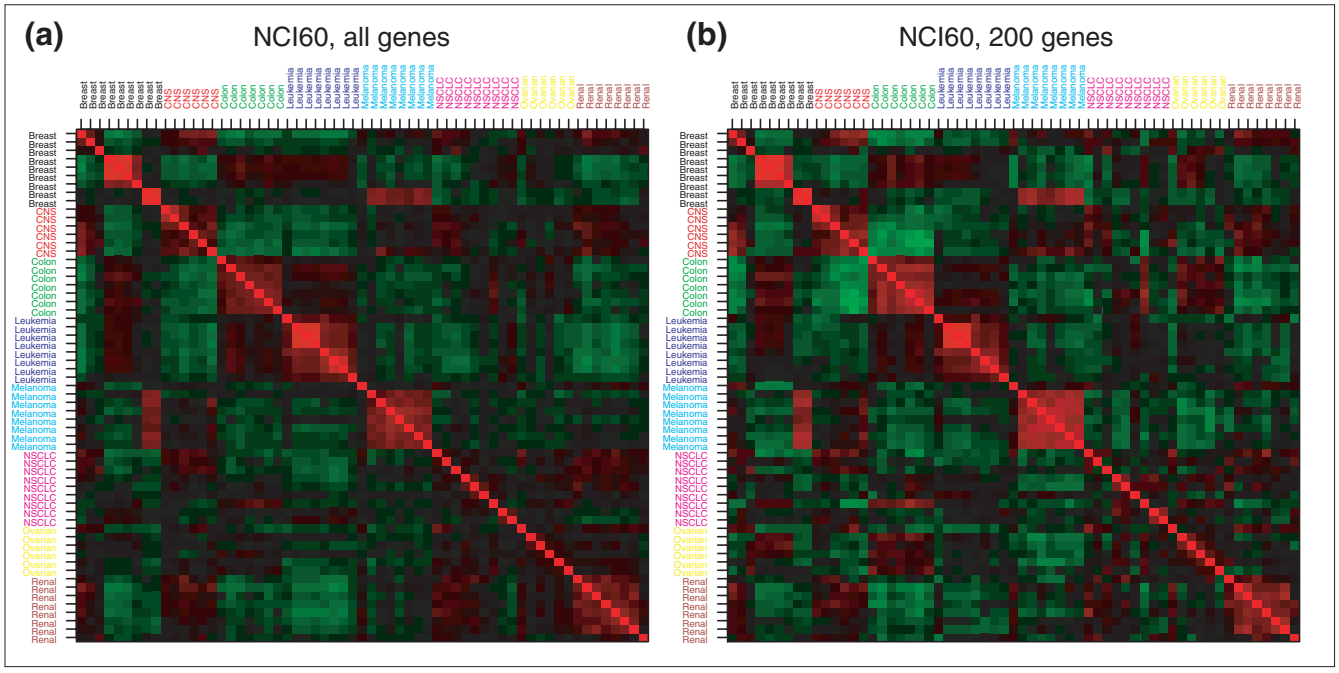

Figure 9
Correlation matrix, NCI60 dataset. Images of the correlation matrix for the 61 cell line mRNA samples based on expression profiles for **(a)** all p = 5,244 genes and **(b)** the p = 200 genes with the largest variance. The mRNA samples are ordered by class: 7 + 2 breast, 6 CNS, 7 colon, 6 + 2 leukemia, 8 melanoma, 9 NSCLC, 6 ovarian, 8 renal.

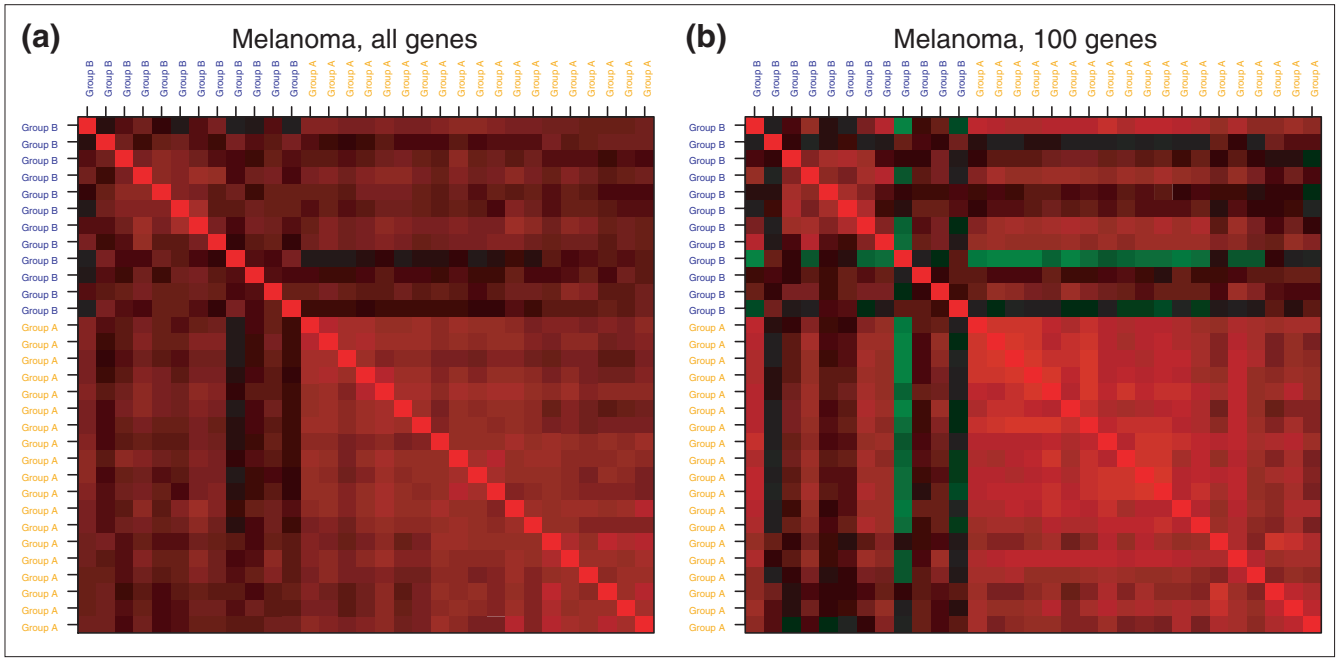

Figure 10
Correlation matrix, melanoma dataset. Images of the correlation matrix for the 31 melanoma mRNA samples based on expression profiles for **(a)** all p = 3,613 genes and **(b)** the p = 100 genes with the largest variance. The mRNA samples are ordered by class, as proposed in [30], first group B (blue), then group A (orange).

DLBCL) are recovered as distinct clusters. For $K = 4$, the largest DLBCL class is divided into two clusters of approximately equal size and the remaining two classes (CLL and FL) are recovered as two distinct clusters. The two DLBCL clusters have a 75% overlap with the subclasses of Alizadeh *et al.* [1]. Finally, for $K = 5$, the smallest class, FL, is divided into two clusters and the rest of the clusters are as with $K = 4$. On the basis of this analysis we do not expect to recover more than four classes in the lymphoma data.

Leukemia

Images of the correlation matrix for the leukemia dataset are displayed in Figure 8. The three classes corresponding to the ALL T-cell, ALL B-cell, and AML samples clearly stand out in the image of the correlation matrix for the 100 genes with the largest variance, but are indistinguishable in the image of the correlation matrix based on all genes. When the PAM procedure is applied to the leukemia dataset using the 100 genes with the largest variance, the results are as follows. For $K = 2$, eight ALL T-cell observations are misclassified with the AML observations. For $K = 3$ classes, one ALL B-cell sample is clustered with the ALL T-cell tumors and the rest of the observations are allocated correctly. For $K = 4$, the ALL B-cell samples are partitioned into two clusters. Finally, for $K = 5$, the AML samples are partitioned into two clusters. On the basis of the correlation matrix, one would expect to identify three tumor classes in this dataset.

NCI60

For the NCI60 cell-line dataset, the classes are not clearly distinguishable from the images of the correlation matrix. Colon, leukemia and melanoma cell lines display the strongest correlations within class, whereas breast, NSCLC and ovarian cell lines seem to be the most heterogeneous classes. When the PAM procedure is applied to the NCI60 dataset using the 200 genes with the largest variance and varying the number of clusters $K \leq 8$, only five types of cell lines tend to cluster together (CNS, colon, leukemia, melanoma, and renal cell lines). On the basis of this observation, one should not expect to recover more than five classes.

Melanoma

Finally, for the melanoma dataset, the image of the correlation matrix for the $p = 100$ most variable genes (Figure 10) could possibly suggest the existence of a subclass of tumors which includes the group A samples of Bittner *et al.* [30]. However, some observations in this cluster (the first one from the left in particular) were not identified by Bittner *et al.* as being part of the tight cluster. Indeed, when PAM is applied to the melanoma dataset using the 100 genes with the largest variance, four additional observations are joined to the 19 observation cluster (group A) proposed by Bittner *et al.* Dividing the data into three clusters results in a split of the 19 observations into two clusters. One would expect to identify, at most, two or three classes for this dataset because of the small sample size.

Acknowledgements

The authors are grateful to Leo Breiman for stimulating discussions on classification. We also acknowledge Terry Speed for discussions on the analysis of microarray data. We thank the referees for their helpful comments and suggestions on an earlier version of this article and for bringing the recent work of Ben-Hur *et al.* to our attention. This work was supported in part by a PMMB Burroughs-Wellcome postdoctoral fellowship (S.D.) and a PMMB Burroughs-Wellcome graduate fellowship (J.F.).

References

1. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, *et al.*: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.** *Nature* 2000, **403:**503-511.
2. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ: **Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.** *Proc Natl Acad Sci USA* 1999, **96:**6745-6750.
3. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh M, Downing JR, Caligiuri MA, *et al.*: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286:**531-537.
4. Perou CM, Jeffrey SS, van de Rijn M, Rees CA, Eisen MB, Ross DT, Pergamenschikov A, Williams CF, Zhu SX, Lee JCF, *et al.*: **Distinctive gene expression patterns in human mammary epithelial cells and breast cancers.** *Proc Natl Acad Sci USA* 1999, **96:**9212-9217.
5. Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, Brown PO: **Genome-wide analysis of DNA copy-number changes using cDNA microarrays.** *Nat Genet* 1999, **23:**41-46.
6. Ross DT, Scherf U, Eisen MB, Perou CM, Spellman P, Iyer V, Jeffrey SS, de Rijn MV, Waltham M, Pergamenschikov A, *et al.*: **Systematic variation in gene expression patterns in human cancer cell lines.** *Nat Genet* 2000, **24:**227-234.
7. Dudoit S, Fridlyand J: *Bagging to Improve the Accuracy of a Clustering Procedure.* Technical Report 600, Department of Statistics, University of California, Berkeley, 2001. [http://www.stat.Berkeley.EDU/tech-reports/index.html]
8. Mardia KV, Kent JT, Bibby JM: *Multivariate Analysis*. San Diego: Academic Press; 1979.
9. Milligan GW: **Clustering validation: results and implications for applied analyses.** In *Clustering and Classification.* Edited by Arabie P, Hubert LJ, Soete GD. River Edge, NJ. World Scientific Publishing; 1996: 341-375.
10. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95:**14863-14868.
11. Fraley C, Raftery A: *How Many Clusters? Which Clustering Method? - Answers via Model-based Cluster Analysis.* Technical Report 329, Department of Statistics, University of Washington, 1998. [http://www.stat.washington.edu/www/research/reports/1990s/#1998].
12. Milligan GW, Cooper MC: **An examination of procedures for determining the number of clusters in a data set.** *Psychometrika* 1985, **50:**159-179.
13. Breiman L: **Bagging predictors.** *Machine Learning* 1996, **24:**123-140.
14. Breiman L: **Arcing classifiers.** *Annls Statistics* 1998, **26:**801-824.
15. Freund Y, Schapire RE: **A decision-theoretic generalization of on-line learning and an application to boosting.** *J Computer System Sci* 1997, **55:**119-139.
16. Kaufman L, Rousseeuw PJ: *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley; 1990.
17. Sarle W: *Cubic Clustering Criterion.* Technical Report A-108, SAS Institute, Inc., 1983. [http://www.sas.com/service/library/onlinedoc/trindex.html].
18. Bock HH: **On some significance tests in cluster analysis.** *J Classification* 1985, **2:**77-108.
19. Hartigan J: **Asymptotic distributions for clustering criteria.** *Annls Statistics* 1978, **6:**117-131.
20. Jain AK, Dubes RC: *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice-Hall; 1988.
21. Calinski R, Harabasz J: **A dendrite method for cluster analysis.** *Commun Statistics* 1974, **3:**1-27.
22. Davies D, Bouldin D: **A cluster separation measure.** *IEEE Trans Pattern Analysis Machine Intelligence* 1979, **1:**224-227.

23. Krzanowski W, Lai Y: **A criterion for determining the number of groups in a dataset using sum of squares clustering.** *Biometrics* 1985, **44:**23-34.
24. Tibshirani R, Walther G, Hastie. T: *Estimating the Number of Clusters in a Dataset via the Gap Statistic.* Technical report, Department of Biostatistics, Stanford University, March 2000. [http://www-stat.stanford.edu/~tibs/research.html]
25. Hartigan JA: **Statistical theory in clustering.** *J Classification* 1985, **2:**63-76.
26. Rand WM: **Objective criteria for the evaluation of clustering methods**. *J Am Stat Assoc* 1971, **66:**846-850.
27. Fowlkes EB, Mallows CL: **A method for comparing two hierarchical clusterings.** *J Am Stat Assoc* 1983, **78:**553-584.
28. Breckenridge J: **Replicating cluster analysis: method, consistency and validity.** *Multivariate Behav Res* 1989, **24:**147-161.
29. Dudoit S, Fridlyand J, Speed TP: **Comparison of discrimination methods for the classification of tumors using gene expression data.** *J Am Stat Assoc* 2002, **97:**77-87.
30. Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, *et al.*: **Molecular classification of cutaneous malignant melanoma by gene expression profiling.** *Nature* 2000, **406:**536-540.
31. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: **Systematic determination of genetic network architecture.** *Nat Genet* 1999, **22:**281-285.
32. McLachlan GJ, Bean R, Peel D: **A mixture model-based approach to the clustering of microarray expression data.** *Bioinformatics* 2002, **18:**413-422.
33. Ben-Hur A, Elisseeff A, Guyon I: **A stability based method for discovering structure in clustered data.** In *Pac Symp Biocomputing* 2002, **7:**6-17.
34. Leisch F: *Bagged Clustering*. Technical report, SFB Adaptive Information Systems and Modelling in Economics and Management Science, Vienna University of Economics and Business Administration, August 1999. [http://www.ci.tuwien.ac.at/~leisch/papers/fl-techrep.html]
35. Yang YH, Buckley MJ, Dudoit S, Speed TP: **Comparison of methods for image analysis on cDNA microarray data.** *J Comput Graph Stat* 2002, **11:**108-136.
36. Yang YH, Dudoit S, Luu P, Speed TP: **Normalization for cDNA microarray data.** In *Microarrays: Optical Technologies and Informatics*. Edited by Bittner ML, Chen Y, Dorsel AN, Dougherty ER. *Proceedings of Society of Photo-Optical Instrumentation Engineers* 2001, **4266:**141-152.
37. **Lymphoma/Leukemia Molecular Profiling Project** [http://genome-www.stanford.edu/lymphoma]
38. **Cancer genomics publications data sets** [http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi]
39. **NCI60 Cancer Microarray Project** [http://genome-www.stanford.edu/nci60]
40. **National Human Genome Research Institute: Resources** [http://www.nhgri.nih.gov/DIR/Microarray/selected_publications.html]
41. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman R: **Missing value estimation methods for DNA microarrays.** *Bioinformatics* 2001, **17:**520-525.