LETTER Communicated by Robert Tibshirani

Stability-Based Validation of Clustering Solutions

Tilman Lange
tilman.lange@info.ethz.ch
Volker Roth
volker.roth@info.ethz.ch
Swiss Federal Institute of Technology (ETH) Zurich, Institute for Computational Science, CH-8092 Zurich, Switzerland

Mikio L. Braun
braunm@cs.uni-bonn.de
Rheinische Friedrich-Wilhelms-Universität Bonn, Institut für Informatik III, 53117 Bonn, Germany

Joachim M. Buhmann
jbuhmann@info.ethz.ch
Swiss Federal Institute of Technology (ETH) Zurich, Institute for Computational Science, CH-8092 Zurich, Switzerland

Data clustering describes a set of frequently employed techniques in exploratory data analysis to extract "natural" group structure in data. Such groupings need to be validated to separate the signal in the data from spurious structure. In this context, finding an appropriate number of clusters is a particularly important model selection question. We introduce a measure of cluster stability to assess the validity of a cluster model. This stability measure quantifies the reproducibility of clustering solutions on a second sample, and it can be interpreted as a classification risk with regard to class labels produced by a clustering algorithm. The preferred number of clusters is determined by minimizing this classification risk as a function of the number of clusters. Convincing results are achieved on simulated as well as gene expression data sets. Comparisons to other methods demonstrate the competitive performance of our method and its suitability as a general validation tool for clustering solutions in real-world problems.

1 Introduction

One of the fundamental tools for understanding and interpreting data when prior knowledge of the underlying distribution is missing is the automatic organization of the data into groups or clusters. Methods for data clustering provide core techniques for exploratory data analysis and have been

successfully applied in numerous domains, such as computer vision, computational biology, and data mining (Jain, Murty, & Flynn, 1999; Gordon, 1999; Buhmann, 1995; Jain & Dubes, 1988). Adopting a machine learning perspective, clustering belongs to the class of unsupervised learning problems. Several steps in a cluster analysis are typically required:

1. The objects that have to be clustered should be represented by informative features. The features often require some preprocessing such as standardization. Furthermore, a suitable similarity measure has to be selected. The choices made here largely predetermine what a clustering algorithm can achieve in the best case.
2. The type of group structure has to be specified in mathematical terms. This choice is essentially determined by selecting a clustering algorithm that encodes a model for the data. Clustering algorithms reflect the structural bias of a grouping principle, for example, that clusters are spherically distributed rather than elongated.
3. The user has to assess the validity of the resulting grouping solution, that is, he has to infer the "correct" number of clusters k as well as the interpretation of the groups. In the absence of prior knowledge, this question represents a particularly hard and important part of the analysis. The main topic of this article is the reliable inference of an appropriate number of clusters or model order.

The notion of cluster validation refers to concepts and methods for the quantitative and objective evaluation of the output of a clustering algorithm (Jain et al., 1999; Gordon, 1999; Jain & Dubes, 1988). A general principle for cluster validation should be applicable to every clustering algorithm and should not be restricted to a specific group of clustering methods. Adopting this point of view, a model validation scheme should avoid relying on additional assumptions about the group structure in the data which have not been captured by the clustering algorithm. Otherwise, it could be easily biased and consequently would lack objectivity.

In this article, such a general principle is proposed by introducing the notion of the stability of clustering solutions. Our approach to cluster validation requires that solutions are similar for two different data sets that have been generated by the same (probabilistic) source. Hence, the replicability of clustering solutions is essentially assessed, an approach also taken by several other methods for cluster validation in general and for estimating the number of clusters in particular. Some methods (Ben-Hur, Elisseeff, & Guyon, 2002; Levine & Domany, 2001) cluster two nondisjoint data sets in order to measure the similarity of the clustering solutions obtained for the intersection of both data sets (see section 3). In section 2, we point out why this approach could be biased. A different approach has been pioneered in an early work by Breckenridge (1989): the basic idea is to measure the agreement of clustering solutions generated by a clustering algorithm and by a

classifier trained using a second (clustered) data set. Breckenridge's work did not lead to a specific implementable procedure, in particular not for model order selection. However, his study suggests the usefulness of such an approach for the purpose of validation. Our method as well as Dudoit and Fridlyand's (2002) Clest, and Tibshirani, Walther, Botstein, and Brown's (2001) Prediction Strength build on Breckenridge's ideas (see sections 2 and 3 for a detailed description of these methods).

The methods described so far can be classified as essentially model free. In contrast to these techniques, model-based approaches (e.g. Smyth, 1998; Fraley & Raftery, 1998) rely on a probabilistic formulation of the clustering problem or implicitly encode assumptions on the type of group structure. Although not explicitly formulated in a probabilistic way, the general validation methodology proposed in Yeung, Haynor, and Ruzzo (2001) as well as the Gap Statistic (Tibshirani, Walther, & Hastie, 2001; see section 3) are model dependent, since both methods assume compact clusters tightly packed around cluster centroids.

Our notion of stability is introduced and developed in section 2 as the expected dissimilarity of clustering solutions that can serve as a confidence measure for data partitioning. Section 3 starts with a short account on the different validation methods used in the experimental study of simulated data sets (see section 3.2) and of gene expression levels measured by DNA microarrays (see section 3.3). The cluster analysis of the data from molecular biology addresses the problem of tumor (sub-)type or class discovery. From the perspective of model order selection, it is necessary to determine how many subtypes can be reliably identified in a given data set.

2 The Stability Measure

The concept of stability as a paradigm to assess solutions to clustering problems is mathematically formalized in this section. Based on this formalism, the number of clusters is estimated from the stability analysis of clustering solutions.

Let $\mathbf{X} = (X_1, \ldots, X_n) \in \mathcal{X}^n$ be a finite data set. The problem of clustering is to find a partition of this data set into k disjoint classes or clusters. A solution is formally represented by a vector of labels $\mathbf{Y} = (Y_1, \ldots, Y_n)$ where $Y_i \in \mathbb{L} := \{1, \ldots, k\}$ and $Y_i = \nu$ if X_i is assigned to cluster ν. A clustering algorithm $\mathcal{A}_k$ constructs such a solution: $\mathbf{Y} := \mathcal{A}_k(\mathbf{X})$.[1] We assume that the data set has been drawn from a probabilistic source. This assumption is particularly reasonable in the context of experimental biological data.

[1] The proposed stability method relies on partitions of the object set as solutions to clustering problems. However, since rooted trees are nested sequences of partitions, the proposed stability approach can be straightforwardly adapted for validating hierarchical groupings.

Strictly speaking, model selection in clustering consists of two steps. First, a grouping algorithm has to be selected. Once this choice is made, the model order remains to be determined for the given data set in the second step. In this contribution, we mainly address the second issue and assume that the user has fixed a clustering algorithm for the application in mind. Thus, in the strict sense, we provide only a partial answer to the problem of cluster model selection. However, due to the lack of a general formal objective in clustering, the question of identifying an appropriate cluster algorithm is ill posed. Posed differently, providing a definitive answer to this question would amount to resolving the clustering problem itself. In spite of this general problem, our experiments indicate that if one grouping strategy exhibits much higher stability than a second one, then the first method is likely to be more appropriate for the data set than the alternative. In case two methods achieve equally high scores, our validation scheme does not prefer one method over the other. This ambiguity, however, cannot be resolved by statistical validation schemes, since the different methods potentially emphasize different aspects of the data. Taking such an abstract viewpoint, we propose that our approach, although primarily designed for solving the model order selection problem, might also help to decide which model is more appropriate in cases where no prior knowledge from the application domain can guide the selection of a clustering algorithm.

The model order selection problem in cluster analysis amounts to finding a suitable number of clusters k. In general, there might exist more than one useful answer to this question. An indispensable requirement, however, is the robustness of a clustering solution, that is, the result of a cluster analysis should be reproducible on other data sets drawn from the same source and should not depend too sensitively on the sample set at hand. This means, for example, that in the context of novel tumor class identification, a second set of tissues originating from the same tumor classes should be clustered in a similar manner. The stability of the clustering solution varies with the number of clusters that are inferred. For instance, inferring too many clusters leads to arbitrary splits of the data, and the solution is influenced heavily by sample fluctuations. Inferring too few clusters might also lead to unstable solutions since the lack of degrees of freedom forces the algorithm to ambiguously mix structures that should be kept separate (the 5 gaussians example in section 3 provides an example for this). Following these considerations, we define the "correct" number of clusters as that of clustering solutions with maximal stability.

The notion of stability that we propose here is based on considering the average dissimilarity of solutions computed on two different data sets. We need to derive a measure of dissimilarity first. Note that a labeling $\mathbf{Y} := \mathcal{A}_k(\mathbf{X})$ is defined only with regard to a data set $\mathbf{X}$. Therefore, solutions $\mathbf{Y} := \mathcal{A}_k(\mathbf{X})$ and $\mathbf{Y}' := \mathcal{A}_k(\mathbf{X}')$ are not directly comparable since they depend on different (often disjoint) data sets $\mathbf{X}$ and $\mathbf{X}'$. For the purpose of assessing the similarity of clustering solutions, a mechanism has to be devised that

renders these solutions comparable. Such a mechanism represents a fundamental difference to the approaches by Levine and Domany (2001) and Ben-Hur et al. (2002), which operate on data sets with overlap.

2.1 Transfer by Prediction. Supervised classification learns a function ϕ that assigns each element of a space $\mathcal{X}$ to one out of k classes based on a labeled input training data set. The function ϕ, which is inferred from the training data set, is called a classifier.

The data set $\mathbf{X}$ together with its clustering solution $\mathbf{Y} := \mathcal{A}_k(\mathbf{X})$ can be considered as a training data set used to construct a classifier. This classifier ϕ trained on $(\mathbf{X}, \mathbf{Y})$ predicts a label $\phi(X')$ for a new object X'. By applying ϕ to each object X'_i in a new set $\mathbf{X}' = (X'_1, \ldots, X'_n)$, we obtain a labeling for $\mathbf{X}'$. For our purpose, we consider the predicted labeling $\phi(\mathbf{X}') := (\phi(X'_i))_{i \le n}$ as the extension of the clustering solution $\mathbf{Y}$ of the data set $\mathbf{X}$ to the data set $\mathbf{X}'$. These predicted labels can be compared to those generated by the clustering algorithm, that is, with $\mathcal{A}_k(\mathbf{X}')$. Hence, the labelings $\mathcal{A}_k(\mathbf{X})$ and $\mathcal{A}_k(\mathbf{X}')$ are turned comparable by utilizing a classifier ϕ as a solution transfer mechanism. The type of classifier largely influences the stability measurement and therefore has to be selected with care, as will be discussed below. After a ϕ has been chosen, the comparison of solutions on the same data set can be addressed.

2.2 Comparing Clustering Solutions. We have derived two solutions $\phi(\mathbf{X}')$ and $\mathbf{Y}'$ for the same data set $\mathbf{X}'$. A very natural distance measure for comparing these two labeling vectors $\phi(\mathbf{X}')$, $\mathbf{Y}'$ is to consider their normalized Hamming distance,

$$d(\phi(\mathbf{X}'), \mathbf{Y}') := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{\phi(X'_i) \neq Y'_i\}, \tag{2.1}$$

where $\mathbf{1}\{\phi(X'_i) \neq Y'_i\} = 1$, if $\phi(X'_i) \neq Y'_i$ and zero otherwise. In the context of supervised classification, this distance between class assignments of an object is called the 0-1 loss function (see, e.g., Duda, Hart, & Stork, 2000). The empirical average over the 0-1 loss of n samples is called empirical misclassification risk. Hence, equation 2.1 can be interpreted as the empirical misclassification risk of ϕ with regard to the test data set $(\mathbf{X}', \mathbf{Y}' = \mathcal{A}_k(\mathbf{X}'))$.

The empirical misclassification risk, equation 2.1, compares two sets of labels that are not necessarily in natural correspondence. Contrary to classification problems, the labeling of the predicted clusters $\phi(X'_i)$ on the second data set has to correspond to the optimized clusters Y'_i only up to an unknown permutation $\pi \in \mathfrak{S}_k$ of the indices, where $\mathfrak{S}_k$ is the set of all permutations of the elements in $\mathbb{L}$. For example, two partitionings of the data set $\mathbf{X}'$ might be structurally equivalent although the labelings $\phi(\mathbf{X}')$ and $\mathbf{Y}'$ are differently represented (i.e., $d(\phi(\mathbf{X}'), \mathbf{Y}') \neq 0$). For example, for $k = 2$, the cluster labeled 1 in the first solution might correspond to the one labeled

2 in the second solution, and vice versa. This ambiguity poses an intrinsic problem in unsupervised learning and arises from the fact that there exists no label information prescribed for the clusters (in contrast to classification).

To overcome the nonuniqueness of representation, we modify the dissimilarity such that the label indices in one solution are optimally permuted to maximize the agreement between the two solutions under comparison, that is,

$$d_{\mathfrak{S}_k}(\phi(\mathbf{X}'), \mathbf{Y}') := \min_{\pi \in \mathfrak{S}_k} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{\pi(\phi(X'_i)) \neq Y'_i\}. \tag{2.2}$$

Technically, the minimization can be performed in time $O(n+k^3)$ by using the Hungarian method (Kuhn, 1955) for minimum weighted bipartite matching, which is guaranteed to find the globally optimal $\pi \in \mathfrak{S}_k$. The procedure requires time $O(n)$ for setting up a weight matrix. The matching itself can be performed in time $O(k^3)$. Thus, the computation of the dissimilarity is still tractable.

So far, a measure of disagreement $d_{\mathfrak{S}_k}$ between solutions generated from two different data sets has been derived. The measure quantifies the fraction of differently labeled points and can be understood as the empirical misclassification risk of ϕ with respect to the "label generator" $\mathcal{A}_k$. Our measure of disagreement is very intuitive since it matches clusters of one solution with those of the other solution in order to maximize the overlap of the respective clusters. Fixing this specific dissimilarity measure represents a significant difference to Clest and the Prediction Strength method (see section 3.1 for a discussion).

2.3 Stability as the Average Distance Between Partitions. Since we consider the data to be drawn from a probability distribution, we are interested in the average distance between solutions where the expectation is taken with regard to pairs of independent data sets $\mathbf{X}, \mathbf{X}'$ of size n, that is,

$$\mathrm{S}(\mathcal{A}_k) := \mathbb{E}_{\mathbf{X},\mathbf{X}'} d_{\mathfrak{S}_k}(\phi(\mathbf{X}'), \mathbf{Y}'). \tag{2.3}$$

We call S the stability index of the clustering algorithm $\mathcal{A}_k$. It is the average dissimilarity of clustering solutions with regard to the distribution of the data. Hence, the smaller the values of the index $\mathrm{S}(\mathcal{A}_k) \in [0, 1]$, the more stable are the solutions. The stability index $\mathrm{S}(\mathcal{A}_k)$ provides a measure of the reproducibility of clustering solutions that are obtained from a probabilistic data source with the clustering algorithm $\mathcal{A}_k$. Conceptually, the expected disagreement between solutions, introduced here as a quality measure, is essentially the quantity of interest in all methods building on the stability idea.

2.4 On the Choice of the Classifier. By selecting an inappropriate classifier, one can artificially increase the discrepancy between solutions, as

Figure 1: Effect of wrong predictor. Using the training data on the left generated with path-based clustering (Fischer, Zöller, & Buhmann, 2001), a nearest centroid classifier (right plot) cannot reproduce the target labeling. In this case, the disagreement between two solutions is artificially increased by the inappropriate predictor. Label indices are represented by symbols in the plots.

illustrated in Figure 1. Choosing a predictor for the stability assessment requires care, a fact disregarded by Dudoit and Fridlyand (2002) as well as by Tibshirani, Walther, Botstein et al. (2001). Since we want to measure the data distribution inherent stability, the influence of the classifier should be minimized, that is, select $\phi^* \in \mathcal{C}$ with minimal misclassification error,

$$\phi^* := \operatorname*{argmin}_{\phi \in \mathcal{C}} \mathbb{E}_{\mathbf{X},\mathbf{X}'} \left[\min_{\pi \in \mathfrak{S}_k} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{\pi(\phi(X'_i)) \neq Y'_i\} \right]. \tag{2.4}$$

Note that the minimization over all predictors is safe and cannot lead to overfitting or trivial stability because ϕ^* minimizes the misclassification on every independent test data set $(\mathbf{X}', \mathbf{Y}')$ for all training data sets $(\mathbf{X}, \mathbf{Y})$ (ϕ^* plays the role here of the Bayes optimal predictor with regard to the label source $\mathcal{A}_k(\mathbf{X})$). Minimizing $\mathrm{S}(\mathcal{A}_k)$ is an important requirement. Consider the predictor that outputs the training labels for objects from the training data set and a constant label (say, 1) to all other, previously unobserved objects. Clearly, minimizing the training error would not suffice since overfitting would remain undetected. Thus, the expected prediction error with regard to the label source $\mathcal{A}_k(\mathbf{X}') = \mathbf{Y}'$ has to be minimized.

Due to the lack of a general finite sample size theory for the accuracy of classifiers, the identification of the optimal classifier by analytical means seems to be unattainable. Therefore, we have to resort to potentially suboptimal classifiers in practical applications. Note, however, that a poorly chosen predictor can only worsen the stability. Equation 2.3 measures the mismatch between clustering and prediction strategy. Intuitively, this mismatch is minimized by mimicking the clustering algorithm. This intuition,

which is confirmed by our experiments, allows us to devise natural, good predictors in practice. For clustering algorithms that minimize a cost function H (as the squared error in k-means clustering or the negative log likelihood in mixture modeling), the grouping strategy can be mimicked if the predictor uses the least-cost increase criterion. Given the training data $(\mathbf{X}, \mathbf{Y})$, assign a new datum X^{new} to the class Y^{new}, for which the cost increase $H((\mathbf{X}, X^{\text{new}}), (\mathbf{Y}, Y^{\text{new}})) - H(\mathbf{X}, \mathbf{Y})$ becomes minimal. The same strategy is applicable for agglomerative algorithms. The working principle of these methods is the least-cost-increase strategy, since they merge the two closest clusters. For k-means, this strategy leads to the nearest centroid classifier up to (negligible) $O(\frac{1}{n})$ corrections. For single linkage, the nearest-neighbor classifier becomes the classifier of choice. In general, there exist algorithms that cannot be easily understood as minimizers of a cost function H (e.g., CLICK by Sharan & Shamir, 2000). For these cases, one can safely resort to K-nearest-neighbor classification to transfer solutions, which is asymptotically Bayes' optimal, at least for metric data.

2.5 Stability for Model Order Selection. To apply our notion of stability to model order selection, some additional considerations are necessary. Note that the range of possible stability values $\mathrm{S}(\mathcal{A}_k)$ depends on the number of clusters k. This dependency arises from the fact that the similarity measure for partitions scales with the number of classes. A misclassification rate of approximately 0.5 for $k = 2$ essentially means that the predictor is as poor as a random predictor. The same misclassification rate, however, for $k = 100$ is far better than random guessing with an asymptotical error of 0.99. Hence, stability indices are not directly comparable for different values of k, but comparability is mandatory to use the stability index for model order selection. The strategy employed here represents one possible way to achieve comparability while keeping the stability measure interpretable. We stress, however, that there does not exist a generally accepted "correct" strategy for normalization and, thereby, scale adjustment.

Note that $\mathrm{S}(\mathcal{A}_k) \leq 1 - \frac{1}{k}$, since for higher-stability index values above the bound, there exists a label permutation π so that the index after permuting drops below this bound (more formally, since $d_{\mathfrak{S}_k}(\phi(\mathbf{X}'), \mathbf{Y}') \leq \mathbb{E}_{\Pi \in \mathfrak{S}_k}[d(\Pi \circ \phi(\mathbf{X}'), \mathbf{Y}')] = 1 - \frac{1}{k}$ assuming a uniform distribution over all $k!$ permutations). The random labeling algorithm $\mathcal{R}_k$ that assigns an object to class ν with probability $\frac{1}{k}$ asymptotically achieves this index value (i.e., for $n \to \infty$). By normalizing the empirical misclassification rate of the clustering algorithm $\mathrm{S}(\mathcal{A}_k)$ with the asymptotic misclassification rate of random labelings $\mathrm{S}(\mathcal{R}_k)$, the stability index values are normalized to the same scale, and thereby comparability is achieved. Hence, we set

$$\bar{\mathrm{S}}(\mathcal{A}_k) := \frac{\mathrm{S}(\mathcal{A}_k)}{\mathrm{S}(\mathcal{R}_k)}. \tag{2.5}$$

$\bar{S}$ is the stability of $\mathcal{A}_k$ relative to the stability achieved by random label guessing.

Note that the stability measure is not defined for $k = 1$. We assume that the data contain nontrivial structure. In our view, the question if the data contain structure (i.e., $k = 1$ or $k > 1$?) should be addressed in advance, for example, by an application-specific test of unimodality. Inclusion of such tests into validity measures can lead to unreliable results, as demonstrated in the experiments. However, the proposed stability index $\bar{S}$ can be considered as a measure of confidence in solutions with $k > 1$. Unusually high instability for $k > 1$ supports the conclusion that only the trivial but uninformative solution (i.e., $k = 1$) contains "reliable" structure.

2.6 Using Stability in Practice. Ideally, one evaluates $\bar{S}$ for several k and chooses the number of clusters k for which a minimum of $\bar{S}$ is obtained. In practical applications, only one data set $\mathbf{X}$ is usually available. However, for evaluating $\bar{S}$, an expected value (with regard to two data sets) has to be estimated using the given data. We propose here a simple subsampling scheme for emulating two independent data sets: iteratively split the data into two halves, and compare the solutions obtained for these halves. Compute estimates for $\bar{S}$ for different k and identify the k^* with the smallest estimate (in case of nonunique minima, the largest k is selected). This k^* is chosen as the estimated number of clusters. A partition of the whole data set can be obtained by applying the clustering algorithm with the chosen k^*. The steps of the procedure are summarized in Table 1.

A few comments are necessary to explain the subsampling scheme. The data should be split into two disjoint subsets because their overlap could otherwise already determine the group structure. This statistical dependence would lead to an undesirable, artificially induced stability. Hence, the use of bootstrapping can be dangerous as well in that it can lead to artificially low

Table 1: Overview of the Algorithm for Stability-Based Model Order Selection.

Repeat for each $k \in \{k_{\min}, \ldots, k_{\max}\}$:

1. Repeat the following steps for r splits of the data to get an estimate $\hat{S}(\mathcal{A}_k)$ by averaging:
- **1a.** Split the given data set into two halves $\mathbf{X}$, $\mathbf{X}'$ and apply $\mathcal{A}_k$ to both.
- **1b.** Use $(\mathbf{X}, \mathcal{A}_k(\mathbf{X}))$ to train the classifier ϕ and compute $\phi(\mathbf{X}')$.
- **1c.** Calculate the distance of the two solutions $\phi(\mathbf{X}')$ and $\mathcal{A}_k(\mathbf{X}')$ for $\mathbf{X}'$ (see eq. (2.2)).

2. Sample s random k-labelings, compare pairs of these, and compute the empirical average of the dissimilarities to estimate $\hat{S}(\mathcal{R}_k)$.

3. Normalize each $\hat{S}(\mathcal{A}_k)$ with $\hat{S}(\mathcal{R}_k)$ to get an estimate for $\bar{S}(\mathcal{A}_k)$.

Return $\hat{k} = \operatorname{argmin}_k \bar{S}(\mathcal{A}_k)$ as the estimated number of clusters.

disagreement between solutions. This is one of the reasons that the methods by Levine and Domany (2001) and Ben-Hur et al. (2002) can produce unreliable results (see section 3). Furthermore, the data sets should have (approximately) equal size so that an algorithm can find similar structure in both data sets. If there are too few samples in one of the two sets, the group structure might no longer be visible for a clustering algorithm. Hence, the option in Clest (see section 3) to have subsets of significantly different sizes can influence the assessment in a negative way and can render the obtained statistics useless in the worst case. Instead of the splitting scheme devised above, one could also fit a generative model to the data from which new (noisified) data sets can be resampled.

3 Experimental Evaluation

In this section, we provide empirical evidence for the usefulness of our approach to model order selection. By using toy data sets, we can study the behavior of the stability measure under well-controlled conditions. By applying our method to gene expression data, we demonstrate the competitive performance of the proposed stability index under real-world conditions. Preliminary results have been presented in Lange et al. (2003).

For the experiments, a deterministic annealing variant of k-means (Rose, Gurewitz, & Fox, 1992) and path-based clustering (Fischer et al., 2001) optimized via an agglomerative heuristic are employed. Averaging is performed over $r = s = 20$ resamples and for $2 \leq k \leq 10$. As discussed in section 2, the predictor should match the grouping principle. Therefore, we use (in conjunction with the stability index) the nearest centroid classifier for k-means and a variant of a nearest-neighbor classifier for path-based clustering that can be considered as a combination of single linkage and pairwise clustering (cf. Fischer et al., 2001; Hofmann & Buhmann, 1997). Concerning the classifier training, see Hastie, Tibshirani, and Friedman (2001).

We empirically compare our method to the Gap Statistic, Clest, Tibshirani's Prediction Strength, Levine and Domany's figure of merit, as well as the model explorer algorithm by Ben-Hur et al. (2002). These methods are described next, with an emphasis on differences and potential shortcomings. The results of this study are summarized in the Tables 2 and 3. Here, $\hat{k}$ is used to denote the estimated number of clusters.

3.1 Competing Methods

3.1.1 The Gap Statistic. Recently, the Gap Statistic has been proposed as a method for estimating the number of clusters (Tibshirani, Walther, & Hastie, 2001). It is not a model-free validation method since it encodes assumptions about the group structure to be extracted. The Gap Statistic relies on considering the total sum of within-class dissimilarities for a given num-

Table 2: Estimated Model Orders for the Toy Data Sets.

Data Set	Stability Method	Gap Statistic	Clest	Prediction Strength	Levine's FOM	Model Explorer	"True" Number k
3 gaussians	$\hat{k} = 3$	$\hat{k} = 3$	$\hat{k} = 3$	$\hat{k} = 3$	$\hat{k} = 2, 3$	$\hat{k} = 2, 3$	$k = 3$
5 gaussians	$\hat{k} = 5$	$\hat{k} = 1$	$\hat{k} = 5$	$\hat{k} = 5$	$\hat{k} = 5$	$\hat{k} = 5$	$k = 5$
3 rings k-means	$\hat{k} = 7$	$\hat{k} = 1$	$\hat{k} = 7$	$\hat{k} = 1$	$\hat{k} = 8$	—	$k = 3$
3 rings path-based	$\hat{k} = 3$	$\hat{k} = 1$	$\hat{k} = 1$	$\hat{k} = 1$	$\hat{k} = 2, 3, 4$	$\hat{k} = 2, 3$	$k = 3$
3 spirals k-means	$\hat{k} = 6$	$\hat{k} = 1$	$\hat{k} = 10$	$\hat{k} = 1$	$\hat{k} = 6$	$\hat{k} = 6$	$k = 3$
3 spirals path-based	$\hat{k} = 3$	$\hat{k} = 1$	$\hat{k} = 6$	$\hat{k} = 1$	$\hat{k} = 2, 3$	$\hat{k} = 3, 6$	$k = 3$

Table 3: Estimated Model Orders for the Gene Expression Data Sets.

Data Set	Stability Method	Gap Statistic	Clest	Prediction Strength	Levine's FOM	Model Explorer	"True" Number k
Golub et al. (1999)	$\hat{k} = 3$	$\hat{k} = 10$	$\hat{k} = 3$	$\hat{k} = 1$	$\hat{k} = 2, 8, 10$	$\hat{k} = 2$	$k \in \{3, 2\}$
Alizadeh et al. (2000)	$\hat{k} = 2$	$\hat{k} = 4$	$\hat{k} = 2$	$\hat{k} = 1$	$\hat{k} = 2, 9$	$\hat{k} = 2$	$k = 3$

ber of clusters k, data set $\mathbf{X}$, and clustering solution $\mathbf{Y} = \mathcal{A}_k(\mathbf{X})$:

$$W_k := \sum_{1 \le \nu \le k} (2n_\nu)^{-1} \sum_{i,j: Y_i = Y_j = \nu} D_{ij}. \tag{3.1}$$

Here, D_{ij} denotes the dissimilarity between X_i and X_j and $n_\nu := |\{i \mid Y_i = \nu\}|$ the number of objects assigned to cluster ν by the labeling $\mathbf{Y}$. If $X_i, X_j \in \mathbb{R}^d$, and $D_{ij} = \|X_i - X_j\|^2$, W_k corresponds to the squared-error criterion optimized by the k-means algorithm. The Gap Statistic investigates the relationship between the $\log(W_k)$ for different values of k and the expectation of $\log(W_k)$ for a suitable null reference distribution through the definition of the gap:

$$\text{gap}_k := \mathbb{E}^*[\log(W_k)] - \log(W_k). \tag{3.2}$$

Here, $\mathbb{E}^*$ denotes the expectation under the null reference distribution (as in Tibshirani, Walther, & Hastie, 2001). A possible reference distribution is the uniform distribution on the smallest hyper-rectangle that contains the

original data. In practice, the expected value is estimated by drawing B samples from the null distribution ($B = 20$ in the experiments), hence

$$\widehat{\text{gap}}_k := \underbrace{\frac{1}{B}\sum_b \log(W^*_{kb})}_{=:\bar{W}^*_k} - \log(W_k), \tag{3.3}$$

where W^*_{kb} is the total within-cluster scatter for the bth data set drawn from the null reference distribution. Now, let std_k be the standard deviation of the sampled $\log(W^*_{kb})$ and $s_k := \text{std}_k\sqrt{1 + 1/B}$. Then the Gap Statistic selects the smallest number of clusters k for which the gap (corrected for the standard deviation) between $\bar{W}^*_k$ and $\log(W_k)$ is large,

$$\hat{k} := \min\{k \mid \widehat{\text{gap}}_k \geq \widehat{\text{gap}}_{k+1} - s_{k+1}\}. \tag{3.4}$$

Since the Gap Statistic relies on W_k in equation 3.1, it presupposes spherically distributed clusters. Hence, it contains a structural bias that should be avoided. The null reference distribution is a free parameter of the method, which essentially determines when to vote for no structure in the data ($\hat{k} = 1$). The experiments in Tibshirani, Walther, and Hastie (2001) as well as the 5 gaussian in section 3.2 demonstrate a sensitivity to the choice of the baseline distribution.

3.1.2 Clest. An approach related to ours has been proposed by Dudoit and Fridlyand (2002). We mainly comment on the differences here. A given data set is repeatedly split into two nondisjoint sets. The sizes of the data sets are free parameters of the method. As we already pointed out in section 2, very unbalanced splitting schemes can lead to unreliable results. After clustering both data sets, a predictor is trained on one data set and tested on the other data set. The predictor is a free parameter of Clest. No guidance is given in Dudoit and Fridlyand (2002) concerning predictor choice. Unfortunately, the subsequent steps are largely determined by the reliability of the prediction step. A similarity measure for partitions is used in order to measure the similarity of the predicted labels to the cluster labels. The measure itself is a free parameter again, in contrast to the method proposed here. In the experiments, the Fowlkes and Mallows (FM) index (Fowlkes & Mallows, 1983) was used.

Given a fixed k, the median t_k of the similarity statistics obtained for B splits of the data is compared to those obtained for B_0 data sets drawn from a null reference distribution. The difference $d_k := t_k - t^0_k$ between the average median statistic under the null hypothesis $t^0_k := (1/B_0)\sum_b t_{k,b}$ and the observed statistic t_k and $p_k := |\{b \mid t_{k,b} \geq t_k\}|B_0^{-1}$ as a measure of variation in the baseline samples are used to select candidate number of

clusters. The number of clusters $\hat{k}$ is chosen by bounding p_k and d_k, yielding two additional free parameters $d_{\min}$ and $p_{\max}$. Dudoit and Fridlyand select

$$\hat{k} = \begin{cases} 1, & \text{if } K^- = \emptyset \\ \min K^-, & \text{otherwise} \end{cases} \tag{3.5}$$

for $K^- := \{2 \leq k \leq M \mid p_k \leq p_{\max}, d_k \geq d_{\min}\}$. The whole procedure is repeated for $k \in \{2, \ldots, M\}$, where M is some predefined upper bound for the number of clusters. Note that the set K^- is essentially determined by the bounds on p_k and d_k. They can be chosen badly so that K^- is always empty, for example.

We conclude that Clest comes with a large number of parameters that have to be set by the user. At the same time, little guidance is given on how to reasonably select values for parameters in practice. This lack of parameter specification poses a severe practical problem since the obtained statistics are of little value for poor parameter selection. In contrast, our method gives guidance concerning the most important parameters. Due to the large degree of freedom in Clest, we consider it a conceptual framework, not a fully specified algorithm. For the experiments, we set $B = B_0 = 20$, $p_{\max} = d_{\min} = 0.05$ and choose a simple linear discriminant analysis classifier as the predictor, following the choice in Dudoit and Fridlyand (2002).

3.1.3 Prediction Strength. This recently proposed method (Tibshirani, Walther, Botstein et al., 2001) is related to ours. Again, the data are repeatedly split into two parts, and the nearest class centroid classifier is employed for prediction. The similarity of clustering solutions is assessed by essentially measuring the intersection of the two clusters in both solutions that match worst. A threshold on the average similarity score is used to estimate the number of clusters. The largest k is selected for which the average similarity is above the user-specified threshold.

From a conceptual point of view as well as in practice, this procedure has two severe drawbacks: (1) it is reasonably applicable to squared-error clustering only due to the use of the nearest centroid predictor, and (2) the similarity measure employed for comparing partitions can trivially drop to zero for large k. In particular, the latter point severely limits the applicability of the Prediction Strength method. In the experiments, we averaged over 20 splits, and the threshold was set to 0.9.

3.1.4 Model Explorer Algorithm. This method (Ben-Hur et al., 2002) clusters nondisjoint data sets: given a data set of size n, two subsamples are generated of size $\lceil fn \rceil$, where $f \in (0.5, 1)$ ($f := 0.8$ in the experiments). The solutions obtained for these subsamples are compared at the intersection of the sets. The similarity measure for partitions is a free parameter of the method (in the experiments, the FM index was used). To estimate k, the experimental section of Ben-Hur et al. (2002) suggests looking for

"a jump in,"

$$\mathbb{P}\{\mathcal{S}_k > \eta\} \approx \frac{1}{r}\sum_{j=1}^{r} \mathbf{1}\{s_{j,k} > \eta\}, \tag{3.6}$$

where $\mathcal{S}_k$ is the similarity score of two k-partitions and $s_{j,k}$ denotes the empirically measured similarity for the jth (out of r) subsample for some fixed η. We choose $\eta = 0.9$ in the experiments.

Looking for a "jump" in the distribution is not a well-defined criterion for determining a number of clusters k as it is qualitative in nature. This vagueness can result in situation where no decision can be made. Furthermore, the Model Explorer algorithm can be biased (toward smaller k) due to the overlapping data sets where the data in both sets can potentially feign stability. As for the other methods, 20 pairs of subsamples were used in the experimental evaluation.

3.1.5 Levine and Domany's Resampling Approach. This method (described here in a simplified way) creates r subsamples of size $\lceil fn \rceil$ where $f \in [0, 1]$ from the original data (Levine & Domany, 2001). For the full data and for the resamples, solutions are computed. They define a figure of merit $\mathcal{M}$ that assesses the average similarity of the solutions obtained on the subsamples with the one obtained on the full sample.

The matrix $T \in \{0, 1\}^{n^2}$ with $T_{ij} := \mathbf{1}\{i \neq j \text{ and } i, j \text{ are in the same cluster}\}$ where $i, j \in \{1, \ldots, n\}$, is called the cluster connectivity matrix, which is a different representation of a clustering solution. The resampling results in a matrix T for the full data and $r \lceil fn \rceil \times \lceil fn \rceil$ matrices $T^{(1)}, \ldots, T^{(r)}$ for the subsamples. For the parameter k, the authors define their figure of merit as

$$\mathcal{M}(k) := \frac{1}{r}\sum_{\rho=1}^{r} \frac{\sum_{i \in J_\rho} \sum_{j \in \mathcal{N}_{\rho,i}} (\delta_{T_{ij}, T_{ij}^{(\rho)}})}{\sum_{i' \in J_\rho} |\mathcal{N}_{\rho,i'}|}, \tag{3.7}$$

where J_ρ is the set of samples in the ρth resample and $\mathcal{N}_{\rho,i}$, $i \in J_\rho$, defines a neighborhood between samples. In the original article, the neighborhood definition was left as a free parameter. The criterion we have chosen for our experiments is called the κ-mutual nearest neighbor neighborhood definition (see Levine, 1999).

$\mathcal{M}(k)$ measures the extent to which the grouping computed on the subsample is in agreement with the solution on the full data set. Thus, $\mathcal{M}(k) = 1$ yields perfect agreement. The authors suggest choosing the parameter(s) k for which local maxima of $\mathcal{M}$ are observed. Several maxima can occur, and hence it is not clear how to choose a single number of clusters in that case. In the experiments, all of them are taken into consideration. We have set $f = 2/3$, $r = 20$, and $\kappa = 20$.

3.2 Experiments Using Toy Data. The first data set consists of three fairly well separated point clouds, generated from three gaussian distributions (25 points from the first and the second and 50 points from the third were drawn) and has also been used by Dudoit and Fridlyand (2002), Tibshirani, Walther, Botstein et al. (2001), and Tibshirani, Walther, and Hastie (2001). For some k, for example, $k = 5$ in Figure 2B, the variance in the stability over different resamples is relatively high. This effect can be explained by the model mismatch: for $k = 5$, the clustering of the three classes depends highly on the subset selected in the resampling. We conclude that additional information about the fit can be obtained from the distribution of the stability index values over the resampled subsets apart from the absolute value of the stability index. For this data set, all methods under comparison are able to infer the "true" number of clusters $k = 3$. Figures 2A and 2B show the clustered data set and the proposed stability index. For $k = 2$, the

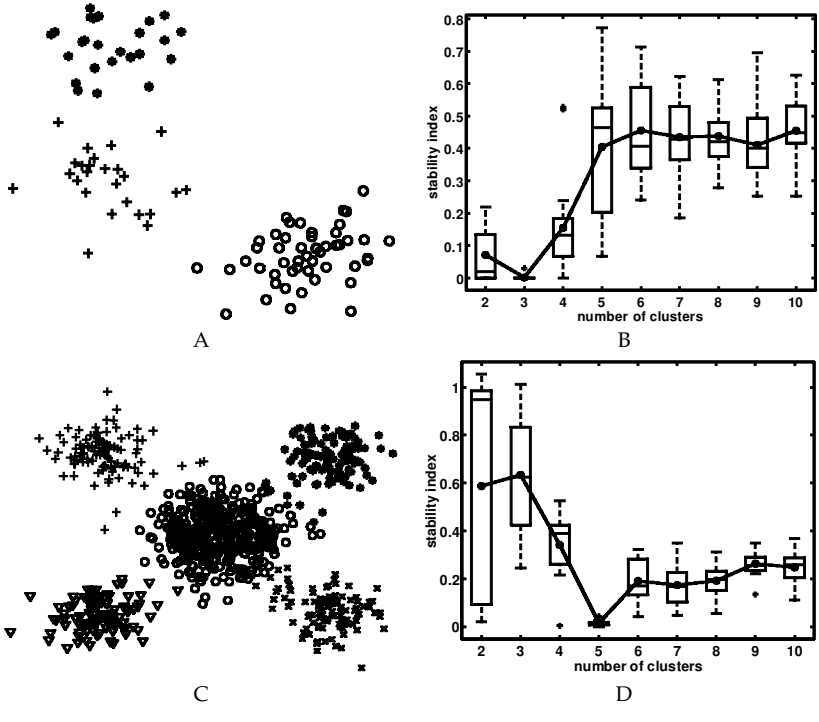

Figure 2: Results of the stability index on the toy data (see section 3.2). (A) Clustering of the 3 gaussians data for $k = 3$. (B) The stability index for the 3 gaussians data set with k-means. (C) Clustering of the 5 gaussians data with the estimated $k = 3$. (D) The stability index for the 5 gaussians data set with k-means.

stability is relatively high due to the hierarchical structure of the data set, which enables stable merging of the two smaller subclusters with 25 data points.

The second proof-of-concept data set consists of $n = 800$ points in $\mathbb{R}^2$ drawn from 5 gaussians (400 from the center one, 100 from the four outer gaussians; see Figure 2C). All methods here estimate the correct number of clusters, $k = 5$, except for the Gap Statistic. In this example the uniformity hypothesis chosen for the Gap Statistic is too strict: groupings of uniform data sets drawn from the baseline distribution have very similar costs to those of the actual data. Hence, there is no large gap for $k > 1$. Figure 2D shows the stability index for this data set. Note the high variability and the poor stability index value for $k < 5$. This high instability is caused by the merging of clusters that do not belong together. Which clusters are merged solely depends on the current noise realization (i.e., on the current subsample).

In the three ring data set (depicted in Figures 3A and 3C), which was also used by Levine and Domany (2001), three ring-shaped clusters can be naturally distinguished. These clusters obviously violate the modeling assumption of k-means of spherically distributed clusters. With $k = 7$, k-means is able to identify the inner circle as a cluster. Thus, the stability for this number of clusters k is highest (see Figure 3B). Clest infers $\hat{k} = 7$, and Levine's FOM suggest $\hat{k} = 8$ while the Gap Statistic and Prediction Strength estimate $\hat{k} = 1$. The Ben-Hur et al. (2002) method does not lead to any interpretable result. Applying the proposed stability estimator with path-based clustering on the same data set yields highest stability for $k = 3$, the "correct" number of clusters (see Figures 3C and 3D). Here, most of the other methods fail and estimate $\hat{k} = 1$. The Gap Statistic fails here because it directly incorporates the assumption of spherically distributed data. Similarly, the Prediction Strength measure and Clest (in the form used here) use classifiers that support only linear decision boundaries, which obviously cannot discriminate between the three ring-shaped clusters. In all these cases, the basic requirement for a validation scheme is violated: it should not incorporate additional assumptions about the group structure in a data set that go beyond the assumptions of the clustering principle employed. Levine's figure of merit as well as Ben-Hur's method infer multiple numbers of clusters. Apart from this observation, it is noteworthy that the stability of k-means is significantly worse than the one achieved with path-based clustering. This stability ranking indicates that the latter is the preferred choice for this data set. Again, the stability can be considered as a confidence measure for solutions. Experiments on uniform data sets, for example, lead to very large values of the stability index (for all $k > 1$), rendering solutions on such data questionable.

The spiral arm data set (shown in Figures 4A and 4C) consists of three clusters formed by three spiral arms. Note that the assumption of compact

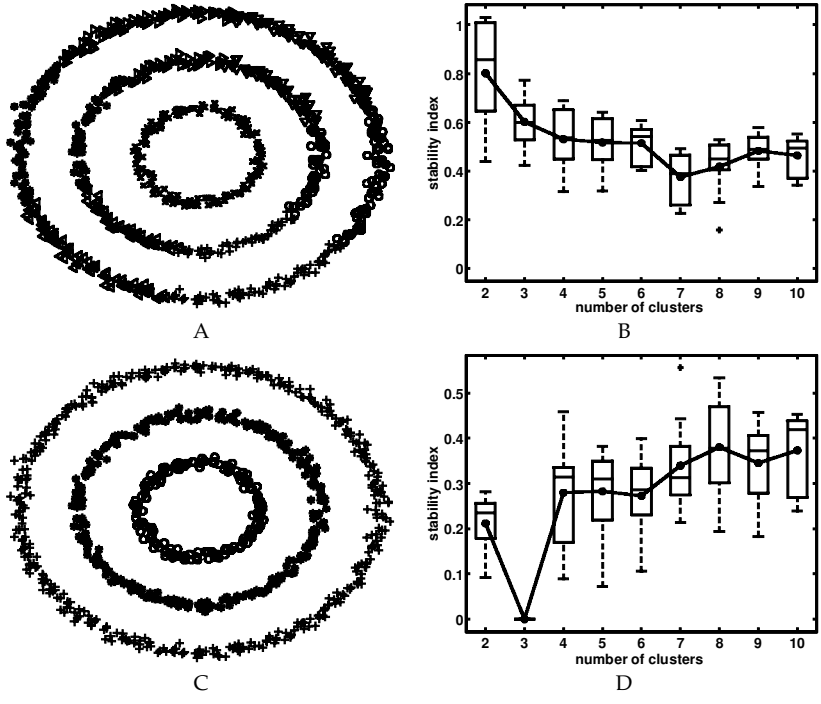

Figure 3: Results of the stability index on the toy data (see section 3.2). (A) k-means clustering solution on the full three-ring data set for $k = 7$. (B) The stability index for the three-ring data set with k-means clustering. (C) Clustering solution on the full data set for $k = 3$ with path-based clustering. (D) The stability index for the three-ring data set with path-based clustering.

clusters is violated again. With k-means, the stability measure overestimates the "true" number of clusters since $\hat{k} = 6$ is returned (see Figure 4B). Most of the other methods fail in this case, since they estimate $\hat{k} = 1$, except for Clest, Levine's FOM, and the Model Explorer algorithm. Clest favors the 10-cluster solution. Note, however, that this $\hat{k}$ is returned because the maximum number of clusters is 10. Levine's FOM and Ben-Hur's method both suggest $k = 6$, as our method does. When path-based clustering is employed, the "correct" number of clusters $k = 3$ is inferred by the stability-based approach (see Figure 4D). In this case, however, most competitors fail again to provide a useful estimate. Only the method by Levine and Domany (2001) and Ben-Hur et al. (2002) estimate $k = 3$ among other number of clusters. Again, the minimum stability index values for path-based clustering are significantly below the index values for k-means. This again indicates that the

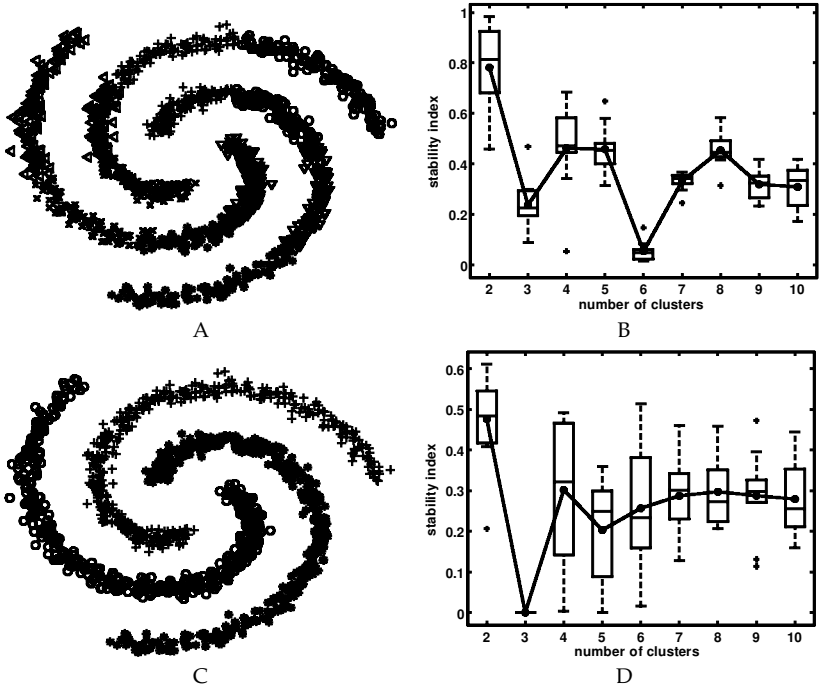

Figure 4: Results of the stability index on the toy data (see section 3.2). (A) Clustering solution on the full spiral arm data set for $k = 6$ with k-means. (B) The stability index for the three-spirals data set with k-means clustering. (C) Clustering solution on the full spiral arm data set for $k = 3$ with path-based clustering (D) and the stability index for this data set.

stability index quantifies the reliability of clustering solutions and, hence, provides useful information for model (order) selection and validation of model assumptions.

3.3 Analysis of Gene Expression Data Sets. With the advent of microarrays, the expression levels of thousands of genes can be simultaneously monitored in routine biomedical experiments (Lander, 1999). Today, the huge amount of data arising from microarray experiments poses a major challenge to gain biological or medical insights. Cluster analysis has turned out to be a useful and widely employed exploratory technique (see, e.g., Tamayo et al., 1999; Eisen, Spellman, Botstein, & Brown, 1998; Shamir & Sharan, 2001) for uncovering natural patterns of gene expression: for example, it is used to find groups of similarly expressed genes. Genes that are clustered together this way are candidates for being coregulated and hence

are likely to be functionally related. Several studies have investigated their data using such an approach (e.g., Iyer et al., 1999).

Another application domain is that of class discovery. Recently, several authors have investigated how novel tumor classes can be identified based exclusively on gene expression data (Golub et al., 1999; Alizadeh et al., 2000; Bittner et al., 2000). A fully Bayesian approach to class discovery is proposed in Roth and Lange (in press). Viewed as a clustering problem, a partition of the arrays is sought that collects samples of the same disease type in one class. Such a partitioning can be used in subsequent steps to identify indicative or explaining genes for the different disease types. The final goal of many such studies is to detect marker genes that can be used to reliably classify and identify new cases. Due to the random nature of expression measurements, the main problem remains to assess and interpret the clustering solutions.

We reinvestigate here the data sets studied by Golub et al. (1999) and Alizadeh et al. (2000) using the stability method to determine a suitable model order. Ground-truth information on the correct number of clusters is available for both data sets that have also been re-analyzed by Dudoit and Fridlyand (2002).

3.3.1 Clustering of Leukemia Samples. Golub et al. (1999) studied in their analysis the problem of classifying acute leukemias. They used self-organizing maps (SOMs; see, Duda et al., 2000) for the study of unsupervised cancer classification. Nevertheless, the important question of inferring an appropriate model order remains unaddressed in their article since a priori knowledge is used to select a number of clusters k. In practice, however, such knowledge is often not available.

Acute leukemias can be roughly divided into two groups, acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). Furthermore, ALL can be subdivided into B-cell ALL and T-cell ALL. Golub et al. (1999) used a data set of 72 leukemia samples (25 AML, 47 ALL of which 38 are B-cell ALL samples).[2] For each sample, gene expression was monitored using Affymetrix expression arrays.

Following the authors of the leukemia study, three preprocessing steps are applied to the expression data. At first, all expression levels are restricted to the interval [100; 16,000], with values outside this interval being set to the boundary values. Second, genes are excluded for which the quotient of maximum and minimum expression levels across samples is smaller than 5 or the difference between maximum and minimum expression is smaller than 500. Finally, the data were $\log_{10}$-transformed. An additional step standardizes samples so that they have zero mean and unit variance across genes. The resulting data set consisted of 3571 genes and 72 samples.

[2] Available on-line at http://www-genome.wi.mit.edu/cancer/.

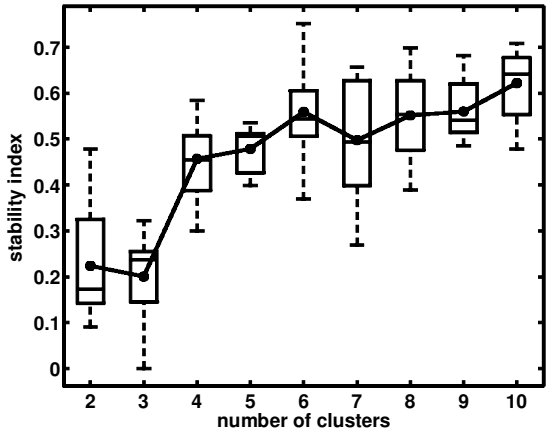

Figure 5: Stability index for the leukemia data set.

For the purpose of cluster analysis, the feature set was additionally reduced by retaining only the 100 genes with highest variance across samples because genes with low variance across samples are unlikely to be informative for the purpose of grouping. This step is adopted from Dudoit and Fridlyand (2002). The final data set consists of 100 genes and 72 samples.

Cluster analysis has been performed with k-means in conjunction with the nearest centroid classifier. Figure 5 shows the stability curve of the analysis for $2 \leq k \leq 10$. For $k = 3$, we estimate the lowest empirical stability index. We expect that clustering with $k = 3$ separates AML, B-cell ALL, and T-cell ALL samples from each other. Figure 6A shows the resulting labeling for this case. With respect to the known ground-truth labels, 91.6% of the samples (66 samples) are correctly classified (the bipartite matching is used again to map the cluster onto the ground-truth labeling). Note that for $k = 2$, similar stability is achieved. Hence, we cluster the data set again for $k = 2$ and compare the result with the ALL – AML labeling of the data. The result is shown in Figure 6A. Here, 86.1% of the samples (62 samples) are correctly identified. The Gap Statistic overestimates the "true" number of clusters, while Prediction Strength does not provide any useful information by returning $\hat{k} = 1$. Clest infers the same number of clusters as our method does. The Model Explorer algorithm reveals a bias to a smaller number of clusters, while Levine's FOM generates several local minima on this data set (not including $k = 3$). We conclude that our method is able to infer biologically relevant model orders. Furthermore, it suggests the model order for which a high accuracy is achieved. If the true labeling had been unknown, our reanalysis of this data set demonstrates that the different cancer classes could have been discovered based exclusively on the expression data by utilizing our model order selection principle and k-means clustering.

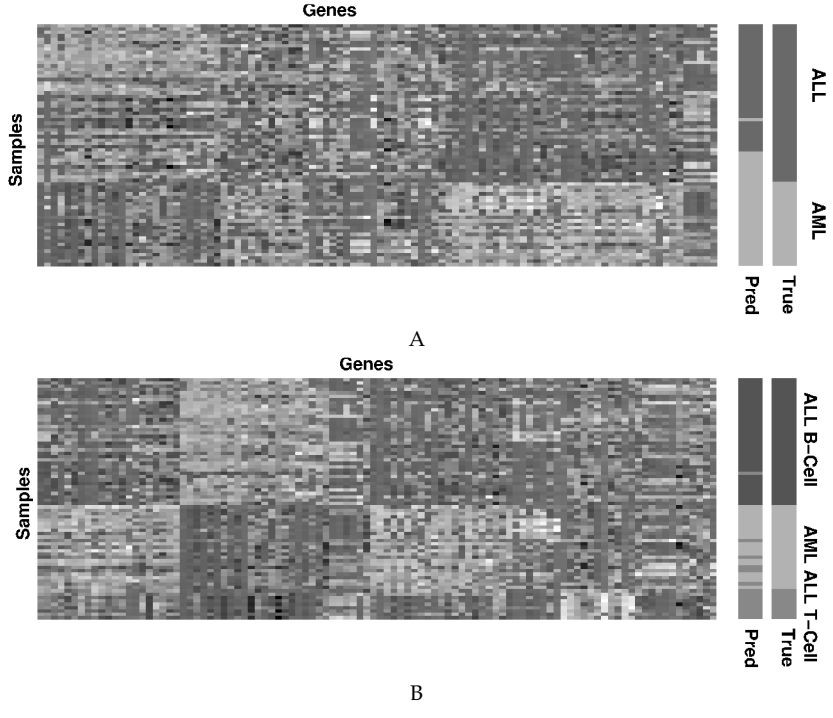

Figure 6: Results of the cluster analysis of the leukemia data set (100 genes, 72 samples) for the most stable model orders $k = 2$ (top) and $k = 3$ (bottom). The data are arranged according to the ground-truth labeling (the vertical bar labeled "True") for both k. The vertical bar "Pred" indicates the predicted cluster labels.

3.3.2 Clustering of Lymphoma Samples. Alizadeh et al. (2000) measured gene expression patterns for three different lymphoid malignancies—diffuse large B-cell lymphoma (DLBCL), follicular lymphoma (FL), and chronic follicular leukemia (CLL)—by utilizing a special microarray chip, the lymphochip. The lymphochip, a cDNA microarray, uses a set of genes that is of importance in the context of lymphoid diseases. The study in Alizadeh et al. (2000) produced measurements for 4682 genes and 62 samples (42 samples of DLBCL, 11 samples of CLL, 9 samples of FL). The data set contained missing values, which were set to 0. After that, the data were standardized as above. Furthermore, the 200 genes with highest variance across samples have been selected for further processing, leading to a data set of 62 samples and 200 genes.

Cluster analysis has been performed with k-means and the nearest centroid rule. The resulting stability curve is depicted in Figure 7. We estimate

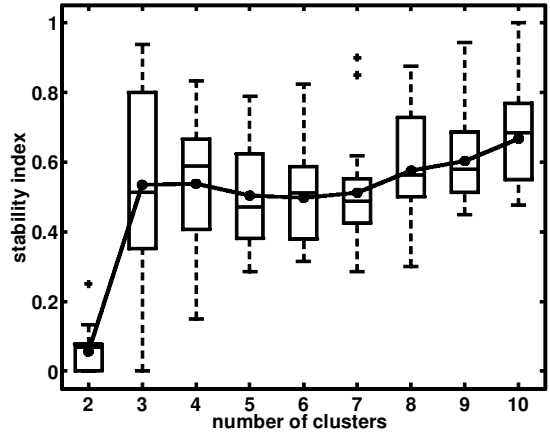

Figure 7: Stability index for the lymphoma data set.

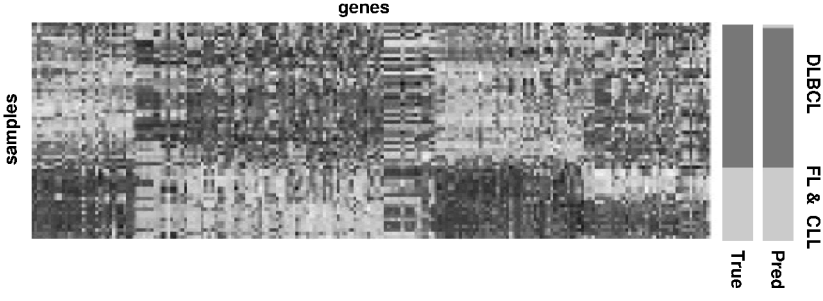

Figure 8: The lymphoma data set extracted from Alizadeh et al. (2000), together with a two-means clustering solution and ground-truth information. k-means almost perfectly separates here DLBCL from CLL and FL. The bar labeled "Pred" indicates the cluster labels, while "true" refers to the ground-truth label information.

here $\hat{k} = 2$, since we get the highest stability index value for this number of clusters. Note that this contradicts the ground-truth number of clusters $k = 3$. Taking a closer look at the solutions (see Figure 8) reveals, however, that the split of the data produced by k-means with $k = 2$ is biologically plausible since it separates DLBCL samples from FL and CLL samples. Furthermore, a 3-means solution does not separate DLBCL, FL, and CLL but splits the DLBCL cluster and generates two clusters consisting of FL, CLL, and DLBCL samples and one cluster of DLBCL samples. In total, we get an agreement of over 98% for $k = 2$, but for $k = 3$, an agreement of only $\approx 57\%$ in comparison with the ground-truth labeling. Hence, our method

estimated the k, for which the most reliable grouping with k-means can be achieved. Clest and Ben-Hur's method achieve the same result as we do. Levine's FOM leads to estimating $k = 2$ and $k = 9$. The Prediction Strength method suggests an uninformative one-cluster solution, while the Gap Statistic proposes $k = 4$. The corresponding grouping solution with k-means again mixes the three different classes in the same clusters. Again, we conclude that we find a biologically plausible partitioning of the data in an unsupervised way. Furthermore, the stability index has selected the number of clusters for which the most consistent data partition with regard to the ground-truth is extracted by k-means clustering.

4 Conclusion

The concept of stability to assess the quality of clustering solutions was introduced. It has been successfully applied to the model order selection problem in unsupervised learning. The stability measure quantifies the expected dissimilarity of two solutions, where one solution was extended by a predictor to the other one. In contrast to other approaches, the important role of the predictor is appreciated in the concept of cluster stability. Under the assumption that the labels produced by the clustering algorithm are the true labels, the disagreement can be interpreted as the attainable misclassification risk for two independent data sets from the same source. Normalizing the stability measure with the stability costs of a random predictor allows us to assess the suitability of different numbers of clusters k for a given data set and clustering algorithm in an objective way. In order to estimate the stability costs in practice, an empirical estimator is used that emulates independent samples by resampling.

Experiments are conducted on simulated and well-known microarray data sets (Alizadeh et al., 2000; Golub et al., 1999). On the toy data, the stability method has demonstrated its competitive performance under well-controlled conditions. Furthermore, we have pointed out where and why competing methods fail or provide unreliable answers. Concerning the gene expression data sets, our reanalysis effectively shows that the stability index is a suitable technique for identifying reliable partitions of the data. The final groupings are in accordance with biological prior knowledge. We conclude that our validation scheme leads to reliable results and is therefore appropriate for assessing the quality of clustering solutions, in particular in biological applications where prior knowledge is rarely available.

Acknowledgments

This work has been supported by the German Research Foundation, grants Buh 914/4, Buh 914/5.

References

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, A., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Judson, J. Jr., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., & Staudt, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature, 403*, 503–511.

Ben-Hur, A., Elisseeff, A., & Guyon, I. (2002). A stability based method for discovering structure in clustered data. In *Pacific Symposium on Biocomputing 2002* (pp. 6–17). Singapore: World Scientific.

Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, S., R., Yakhini, Z., Ben-Dor, A., Sampas, N., Dougherty, E., Wang, E., Marincola, F., Gooden, C., Lueders, J., Glatfelter, A., Pollock, P., Carpten, J., Gillanders, E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D., & Sondak, V. (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature, 406*(3), 536–540.

Breckenridge, J. (1989). Replicating cluster analysis: Method, consistency and validity. *Multivariate Behavioral Research, 24*, 147–161.

Buhmann, J. M. (1995). Data clustering and learning. In M. Arbib (Ed.), *Handbook of brain theory and neural networks* (pp. 278–282). Cambridge, MA: MIT Press.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification.* New York: Wiley.

Dudoit, S., & Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology, 3*(7). Available on-line: http://genomebiology.com/2002/317/research/0036.

Eisen, M., Spellman, P., Botstein, D., & Brown, P. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA, 95*, 14863–14868.

Fischer, B., Zöller, T., & Buhmann, J. M. (2001). Path based pairwise data clustering with application to texture segmentation. In M. A. T. Figueiredo, J. Zerubia, & A. K. Jain (Eds.), *LNCS energy minimization methods in computer vision and pattern recognition.* Berlin: Springer-Verlag.

Fowlkes, E., & Mallows, C. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association, 78*, 553–584.

Fraley, C., & Raftery, A. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer Journal, 41*(8), 578–588.

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

Gordon, A. D. (1999). *Classification* (2nd ed.). London: Chapman & Hall.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference and prediction.* New York: Springer-Verlag.

Hofmann, T., & Buhmann, J. M. (1997). Pairwise data clustering by deterministic annealing. *IEEE PAMI, 19*(1), 1–14.

Iyer, V. R., Eisen, M. B., Ross, D. T., Schuler, G., Moore, T., Lee, J. C. F., Trent, J. M., Staudt, L. M., Hudson, J. Jr., Boguski, M. S., Lashkari, D., Shalon, D., Botstein, D., & Brown, P. O. (1999). The transcriptional program in the response of human fibroblasts to serum. *Science, 283*(1), 83–87.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data.* Upper Saddle River, NJ: Prentice Hall.

Jain, A. K., Murty, M., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys, 31*(3), 265–323.

Kuhn, H. (1955). The Hungarian method for the assignment problem. *Naval Res. Logist. Quart., 2*, 83–97.

Lander, E. (1999). Array of hope. *Nature Genetics Supplement, 21*, 3–4.

Lange, T., Braun, M., Roth, V., & Buhmann, J. (2003). Stability-based model selection. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems, 15* (pp. 617–624). Cambridge, MA: MIT Press.

Levine, E. (1999). *Un-supervised estimation of cluster validity—methods and applications.* Master's thesis, Weizmann Institute of Science.

Levine, E., & Domany, E. (2001). Resampling method for unsupervised estimation of cluster validity. *Neural Computation, 13*, 2573–2593.

Rose, K., Gurewitz, E., & Fox, G. (1992). Vector quantization and deterministic annealing. *IEEE Trans. Inform. Theory, 38*(4), 1249–1257.

Roth, V., & Lange, T. (in press). Bayesian class discovery in microarray datasets. *IEEE Transactions on Biomedical Engineering.*

Shamir, R., & Sharan, R. (2001). Algorithmic approaches to clustering gene expression data. In T. Jiang, T. Smith, Y. Xu, & M. Q. Zhang (Eds.), *Current topics in computational biology.* Cambridge, MA: MIT Press.

Sharan, R., & Shamir, R. (2000). CLICK: A clustering algorithm with applications to gene expression analysis. In *ISMB'00* (pp. 307–316). Menlo Park, CA: AAAI Press.

Smyth, P. (1998). *Model selection for probabilistic clustering using cross-validated likelihood* (Tech. Rep. 98-09). Irvine, CA: Information and Computer Science, University of California, Irvine.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., & Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *PNAS, 96*, 2907–2912.

Tibshirani, R., Walther, G., Botstein, D., & Brown, P. (2001). *Cluster validation by prediction strength* (Tech. Rep.). Stanford, CA: Statistics Department, Stanford University.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters via the gap statistic. *J. Royal Statist. Soc. B, 63*(2), 411–423.

Yeung, K., Haynor, D., & Ruzzo, W. (2001). Validating clustering for gene expression data. *Bioinformatics, 17*(4), 309–316.

Received May 15, 2003; accepted November 24, 2003.