Bayesian Class Discovery in Microarray Datasets

Volker Roth* and Tilman Lange, *Student Member, IEEE*

***Abstract*—A novel approach to class discovery in gene expression datasets is presented. In the context of clinical diagnosis, the central goal of class discovery algorithms is to simultaneously find putative (sub-)types of diseases and to identify informative subsets of genes with disease-type specific expression profile. Contrary to many other approaches in the literature, the method presented implements a *wrapper* strategy for feature selection, in the sense that the features are directly selected by optimizing the discriminative power of the used partitioning algorithm. The usual combinatorial problems associated with wrapper approaches are overcome by a Bayesian inference mechanism. On the technical side, we present an efficient optimization algorithm with guaranteed local convergence property. The only free parameter of the optimization method is selected by a resampling-based stability analysis. Experiments with Leukemia and Lymphoma datasets demonstrate that our method is able to correctly infer partitions and corresponding subsets of genes which both are relevant in a biological sense. Moreover, the frequently observed problem of ambiguities caused by different but equally high-scoring partitions is successfully overcome by the model selection method proposed.**

***Index Terms*—Automatic relevance determination, Bayesian inference, class discovery, gene expression, gene selection.**

I. Introduction and Related Work

A central goal of the analysis of microarray data is the identification of small subsets of informative genes with disease-specific expression profiles. While the problem of selecting genes for *known* disease types has been studied widely in the literature, the *discovery* of putative *sub-types* of diseases is still a challenging task. Taking a machine-learning viewpoint, this *class-discovery* problem can be formalized as an unsupervised clustering problem with simultaneous feature selection. Early approaches to this problem [1]–[3], were semi-automatic procedures based on a combination of clustering techniques and human intervention for selecting "relevant" genes. Several shortcomings of such approaches, and also some methods for overcoming these problems, have been discussed in the literature, e.g., [4]–[6]. The common strategy of most of these approaches is the use of a (possibly iterated) stepwise procedure, in which the first step consists of extracting a set of hypothetical partitions (the *clustering* step), and the other step involves some way of scoring genes for relevance (the *relevance determination* step). A possible shortcoming of these approaches is the way of combining these two steps in an "*ad hoc*" manner: usually the relevance determination mechanism does not take into account the properties of the clustering method used. It rather attempts to find predictive subsets of genes by making use of simple empirical statistical measures, such as T-test scores or correlation coefficients. These scoring measures treat the genes as independent objects, while ignoring both the biological knowledge of gene expression levels being correlated, and also ignoring the inherent capabilities of many clustering methods for handling such correlations.

In *supervised* learning scenarios, feature selection methods of this kind are called *filter methods*, whereas the so called *wrapper methods* directly make use of the classification algorithm. From a conceptual viewpoint, wrapper approaches are clearly advantageous, since the features are selected by optimizing the discriminative power of the finally used classifier. Returning to our *unsupervised* class-discovery scenario, it is thus of particular interest to incorporate wrapper strategies for gene selection into clustering methods. The approach to class discovery that we present in this paper, can be viewed as a method of this kind. It combines a Gaussian mixture model with a Bayesian feature selection principle. Features are selected by maximizing a constrained likelihood criterion, without making limiting assumptions of independence. The usual combinatorial problems involved with wrapper approaches are overcome by using a Bayesian inference mechanism for selecting the relevant features. We present an efficient optimization algorithm for our model with guaranteed convergence to a local optimum. The only free parameter of the optimization method is selected by a resampling-based stability analysis.[1] Experiments for real-world datasets demonstrate that this model selection mechanism is capable of selecting stable and reliable partitions. On the one hand, this mechanism overcomes the problem of many ambiguous and equally high-scoring splitting hypotheses, which seems to be an inherent shortcoming of many approaches that have been proposed in the literature. On the other hand, a comparison with ground-truth labels in control experiments indicate that the selected models lead to partitions which have a clear biological meaning.

The remainder of this paper is organized as follows. Section II presents the theoretical derivation of the proposed class discovery model. For a Gaussian mixture model we first introduce the expectation-maximization (EM) algorithm, in which the M-step has been replaced by a linear discriminant analysis (LDA). Then, we incorporate a Bayesian feature selection mechanism into this LDA-based M-step. The final EM algorithm has only one free model parameter, for which we propose a stability-based selection strategy in Section III. The main focus of Sections IV and V concerns the application of the method proposed for two cancer datasets.

Manuscript received April 8, 2003; revised August 10, 2003. This work was supported in part by the DFG under Grant BU 914/4 and Grant BU 914/5. *Asterisk indicates corresponding author.*

*V. Roth is with Institute for Computational Science, ETH Zurich, Hirschengraben 84, CH-8092 Zurich, Switzerland (e-mail: vroth@inf.ethz.ch).

T. Lange is with the Institute for Computational Science, ETH Zurich, CH-8092 Zurich, Switzerland.

Digital Object Identifier 10.1109/TBME.2004.824139

[1]The whole processing pipeline of a real world experiment may include additional parameters related to data preprocessing and/or model selection, see Section IV for details.

II. Class Discovery and Bayesian Relevance Determination

A. Technical Overview

Our approach to class discovery is based on a Gaussian mixture model, which is optimized by way of the classical *expectation-maximization* (EM) algorithm. In order to incorporate the feature selection mechanism, the maximization (M)-step of this algorithm is first reformulated as a linear discriminant analysis problem, which in turn is carried out by optimizing a linear regression functional. We then take a Bayesian perspective and specify a prior distribution over the regression coefficients, which has the functional form of a so called *Automatic Relevance Determination* prior. For each regression coefficient, this prior contains a free hyperparameter, which encodes the "relevance" of the corresponding variable in the linear regression. In a Bayesian inference step, these hyperparameters are then integrated out from the distribution over the regression coefficients. We finally arrive at a M-step with integrated feature selection principle. After iterated application of E- and M-step, the algorithm is proven to converge to a local optimum.

B. Gaussian Mixtures and LDA

For the task of class discovery in gene expression experiments, the data is given as a collection of N microarrays (in the following called *samples*), each of which contains expression levels of d' genes. Usually, microarray experiments involve some sort of data preprocessing, like standardization and preselection of a subset of $d \leq d'$ genes with high expression variances across the different microarrays. We can thus formalize the input data as a set of d-dimensional vectors $\{\boldsymbol{x}_i\}_{i=1}^N$, $\boldsymbol{x}_i \in \mathbb{R}^d$.

For the purpose of finding *sample clusters*, consider now a Gaussian mixture model with k mixture components which share an identical covariance matrix Σ. Under this model, the log-likelihood for the dataset $\{\boldsymbol{x}_i\}_{i=1}^N$ reads

$$l^{mix} = \sum_{i=1}^{N} \log \left(\sum_{\nu=1}^{k} \pi_\nu \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_\nu, \Sigma) \right) \tag{1}$$

where the mixing proportions π_ν sum to one, and ϕ denotes a Gaussian density. The classical EM-algorithm, [7], provides a convenient method for maximizing l^{mix}

E-step: set

$$p_{\eta i} = \text{Prob}(\boldsymbol{x}_i \in \text{class } \eta) = \frac{\pi_\eta \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_\eta, \Sigma)}{\sum_{\nu=1}^{k} \pi_\nu \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_\nu, \Sigma)}.$$

M-step: set

$$\boldsymbol{\mu}_\nu = \frac{\sum_{i=1}^{N} p_{\nu i} \boldsymbol{x}_i}{\sum_{i=1}^{N} p_{\nu i}}$$

$$\Sigma = \frac{1}{N} \sum_{\nu=1}^{k} \sum_{i=1}^{N} p_{\nu i} (\boldsymbol{x}_i - \boldsymbol{\mu}_\nu)(\boldsymbol{x}_i - \boldsymbol{\mu}_\nu)^T$$

The likelihood equations in the M-step can be viewed as weighted mean and covariance maximum likelihood estimates in a weighted and augmented problem: one replicates the N observations k times, with the ν-th such replication having observation weights $p_{\nu i}$. In [8] it is proven that the M-step can be carried out via a weighted and augmented *linear discriminant analysis* (LDA). Following [9], any LDA problem can be restated as an *optimal scoring* problem. Let the class-memberships of the N data vectors be represented by a categorical response variable $\mathcal{G}$ with k levels. Let these responses be coded as a matrix Z, the i, ν-th entry of which equals one if the i-th observation belongs to class ν. The point of optimal scoring is to turn categorical variables into quantitative ones: the score vector $\boldsymbol{\theta}$ assigns the real number θ_ν to the ν-th level of $\mathcal{G}$. The simultaneous estimation of scores and regression coefficients $\boldsymbol{\beta}$ constitutes the optimal scoring problem: minimize the squared residual

$$M(\boldsymbol{\theta}, \boldsymbol{\beta}) = \|Z\boldsymbol{\theta} - X\boldsymbol{\beta}\|_2^2 \tag{2}$$

under the constraint $(1/N)\|Z\boldsymbol{\theta}\|_2^2 = 1$. The notion $\|\cdot\|_2^2$ stands for the squared ℓ_2-norm, and X denotes the (centered) data matrix, the rows of which consist of the input vectors $\boldsymbol{x}_i$. In [9] an algorithm for carrying out this optimization has been proposed, the main ingredient of which is a linear multi-response regression of the data matrix X against the scored indicator matrix $Z\boldsymbol{\theta}$.

Returning from a general LDA problem to the above weighted and augmented problem, it turns out that it is *not* necessary to explicitly replicate the observations: the optimal scoring version of LDA provides an implicit solution of the augmented problem that still uses only N observations. Instead of using a response indicator matrix Z, one uses a *blurred* response matrix $\tilde{Z}$, whose rows consist of the current class probabilities for each observation. At each M-step this $\tilde{Z}$ is used in a linear regression, see [8].

After iterated application of the E- and M-step, an observation $\boldsymbol{x}_i$ is finally assigned to the class ν with highest probability of membership $p_{\nu i}$. Note that the EM iterations are guaranteed to converge monotonically to a local maximum of the likelihood.

C. LDA and Automatic Relevance Determination

We now focus on incorporating the automatic feature selection mechanism into the EM-algorithm. To guarantee a clear theoretical interpretation, however, it will be necessary to restrict the general problem with k mixture components to a 2-component problem. For handling multiple groups, we propose to use the 2-component algorithm in a hierarchical manner by iteratively splitting the clusters into two subclusters. Despite the potential shortcomings of such an iterative splitting approach, our experiments suggest that this hierarchical splitting works very well in practical applications.

According to [9], a 2-class LDA problem can be solved by the following algorithm.

1) Choose an initial N-vector of scores $\boldsymbol{\theta}_0$ which satisfies the constraint
$$N^{-1}\boldsymbol{\theta}_0^T \tilde{Z}^T \tilde{Z} \boldsymbol{\theta}_0 = 1$$
and is orthogonal to a 2-vector of ones $(1,1)^T$. Set $\boldsymbol{\theta}^* = \tilde{Z}\boldsymbol{\theta}_0$;
2) Run a linear regression of X on $\boldsymbol{\theta}^*$: $\hat{\boldsymbol{\theta}}^* = X(X^TX)^{-1}X^T\boldsymbol{\theta}^* =: X\boldsymbol{\beta}$.

The feature selection mechanism can now be incorporated in the M-step by imposing a certain constraint on the linear regression [step 2) of the above algorithm]. In [9], [10] it has been proposed to use a ridge-type penalized regression. On the technical side, such a penalized regression model is obtained by substituting the covariance matrix $(1/N)(X^TX)$ by a penalized version of the form $(1/N)(X^TX + \lambda I)$. In such a ridge regression model, the parameter λ has the role of the Lagrange parameter in a ℓ_2-constrained optimization problem: minimize the functional (2) subject to $\sum_{i=1}^{d} \beta_i^2 = ||\boldsymbol{\beta}||_2^2 < \kappa$. The main idea of incorporating an automatic feature selection mechanism consists of replacing the ℓ_2-penalty by an ℓ_1-penalty: minimize (2) subject to $\sum_{i=1}^{d} |\beta_i| = ||\boldsymbol{\beta}||_1 < \kappa$. In the statistical literature, this model is known as the *Least Absolute Shrinkage and Selection Operator* (LASSO), [11]. In [12], [13] it has been shown that the LASSO model can be interpreted as a Bayesian inference mechanism for the following model: consider *automatic relevance determination* (ARD) priors over the regression coefficients[2]:

$$p(\boldsymbol{\beta}|\boldsymbol{\vartheta}) = \prod_i \mathcal{N}\left(0, \vartheta_i^{-1}\right) \propto \exp\left[-\sum_i \vartheta_i \beta_i^2\right]. \quad (3)$$

In this case, each coefficient β_i has its own prior variance ϑ_i^{-1}. Note that in the above ARD framework only the functional form of the prior (3) is fixed, whereas the parameters ϑ_i, which encode the "relevance" of each variable, are estimated from the data. In [15] the following Bayesian inference procedure for the prior parameters has been introduced: given exponential hyperpriors, (the variances ϑ_i^{-1} must be nonnegative), $p(\vartheta_i) = (\gamma/2)\exp\{-(\gamma\vartheta_i/2)\}$, one can analytically integrate out the hyperparameters from the prior distribution over the coefficients β_i:

$$p(\beta_i) = \int_0^\infty p(\beta_i|\vartheta_i)p(\vartheta_i)d\vartheta_i = \frac{\gamma}{2}\exp\left\{-\sqrt{\gamma}|\beta_i|\right\}. \quad (4)$$

Switching to the maximum a posteriori (MAP) solution in log-space, this marginalization directly leads us to the above ℓ_1-constrained LASSO problem:

$$M(\boldsymbol{\theta}, \boldsymbol{\beta})_{\text{LASSO}} = ||Z\boldsymbol{\theta} - X\boldsymbol{\beta}||_2^2 + \tilde{\lambda}||\boldsymbol{\beta}||_1, \quad (5)$$

where we have defined the Lagrange parameter $\tilde{\lambda} := \sqrt{\gamma}$.

Returning to (3), we are now able to interpret the LASSO estimate as a Bayesian feature selection principle: for the purpose of feature selection, we would like to estimate the value of a binary selection variable $\mathcal{S}$ for each feature: $\mathcal{S}_i$ equals one, if the i-th feature is considered relevant for the given task, and zero otherwise. Taking into account feature correlations, estimation of $\mathcal{S}_i$ necessarily involves searching the space of all possible subsets of features containing the i-th one. In the Bayesian ARD formalism, this combinatorial explosion of the search space is overcome by relaxing a binary selection variable to a positive real-valued variance of a Gaussian prior over each component of the coefficient vector. Following the Bayesian inference principle, we introduce hyperpriors and integrate out these variances, and we finally arrive at the ℓ_1-constrained LASSO problem. During optimization of the LASSO functional, it turns out that many coefficients β_i are shrinked to zero, and the corresponding features are removed from the model.

Summing up: The EM-algorithm with incorporated feature selection, is shown in the equation at the bottom of the page.

D. Optimizing the Final Model

Since space here precludes a detailed discussion of ℓ_1-constrained regression problems, the reader is referred to [16], where a highly efficient algorithm with guaranteed global convergence has been proposed. For our iterated EM-model we can guarantee convergence to a local maximum of the constrained likelihood. Consider two cases: 1) the unconstrained solution is feasible. In this case our algorithm simply reduces to the standard EM procedure, for which it is known that in every iteration the likelihood monotonically increases; 2) the ℓ_1-constraint is active. Then, in every iteration the LASSO algorithm maximizes the likelihood within the feasible region of β-values defined by $||\boldsymbol{\beta}||_1 < \kappa$. The likelihood cannot be decreased in further stages of the iteration, since any solution $\hat{\boldsymbol{\beta}}$ found in a preceding iteration is also a valid solution for the actual problem (note that κ is fixed!). In this case, the algorithm has converged to a local maximum of the likelihood within the constraint region.

III. Model Selection

Apart from the "core" model for clustering and feature selection, a real-world class discovery experiment involves several additional steps, such as data preprocessing, adjustment of model parameters and interpretation of the results. While issues on data preprocessing will be addressed in Section IV-A, here

[2]For an introduction to the ARD principle the reader is referred to [14].

$$\begin{array}{cc}
\mathbf{E-step} & \mathbf{M-step} \\
\downarrow & \text{M}-\text{step} \downarrow \text{carried out by LDA} \\
\downarrow & \text{LDA as} \downarrow \text{linear regression} \\
\downarrow & \text{ARD priors} \downarrow p(\boldsymbol{\beta}|\boldsymbol{\vartheta}) = \prod_i \mathcal{N}(0, \vartheta_i^{-1}) \\
\downarrow & \text{Hyperpriors} \downarrow p(\vartheta_i) = \frac{\gamma}{2}\exp\left\{-\frac{\gamma\vartheta_i}{2}\right\} \\
\downarrow & \text{Marginalization} \downarrow p(\beta_i) = \int_0^\infty p(\beta_i|\vartheta_i)p(\vartheta_i)d\vartheta \\
\mathbf{Estimate\ prob}(\boldsymbol{x}_i \in \text{class}\ \eta) & \mathbf{Optimize\ LASSO}
\end{array}$$

we focus on selecting the value of the ℓ_1-constraint κ in the generalized EM-algorithm described in the last section.

In particular, this section focuses on a method for selecting κ by observing the *stability* of data partitions. For each of the partitions we have identified as "stable," we then examine the fluctuations involved in the feature selection process. It should be noticed that the concept of measuring the stability of solutions as a means of model selection and model assessment has been successfully applied to several unsupervised learning problems, see, e.g. [17]–[19].

Our clustering method splits the data in two disjoint groups, and simultaneously selects features (i.e. prototypical expression patterns of gene-clusters) which support the splitting hypothesis. In a large microarray dataset we will usually find many potential splits, depending on how many features are selected: if we select only a very small number of features (say one), it is likely to find many competing hypotheses for splits. The problem is that usually all single features (i.e. all single genes) will vote for a different sample partition. Taking into account the problem of noisy measurements in microarray experiments, the finally chosen partition will probably tell us more about the exact noise realization than about meaningful splits. If, on the other hand, we select too many features, we face the usual problems of finding structure in very high-dimensional datasets: our functional which we want to optimize will have many local minima, and the optimization algorithm will pick one by chance. Moreover, it is likely that there exists no distinct global optimum, since the expression patterns on the whole chip will vote for contradictory hypotheses. Between these two extremes, we can hope to find relatively stable splits, which are robust against noise, and which are also robust against inherent instabilities of the optimization procedure.

In order to obtain a quantitative measure of stability, we propose the following procedure: run the class discovery method once, corrupt the expression levels by a small amount of noise, repeat the grouping procedure, and calculate the Hamming distance between the two partitions as a measure of (in-)stability. For computing Hamming distances, the partitions are viewed as vectors containing the cluster labels. Simply taking the average stability over many such two-sample comparisons, however, would not allow an adequate handling of situations where there are two equally likely stable solutions, of which the clustering algorithm randomly selects one. In such situations, the averaged stability will be very low, despite the fact that there are two stable splitting hypotheses. This problem can be overcome by looking for compact *clusters* of highly similar partitions, leading to the following refined algorithm:

Algorithm for identifying stable partitions: for different values of the ℓ_1-constraint κ do

(i) compute m noisy replications of the data
(ii) run the class discovery algorithm for each of these datasets
(iii) compute the $m \times m$ matrix of pairwise Hamming distances between all partitions
(iv) cluster the partitions into compact groups and score the groups by their relative frequency
(v) select dominant groups of partitions and choose representants

In step (i) a "suitable" noise level must be chosen *a priori*. In Section IV-A we will discuss how to select the amount of noise by analyzing the typical variations within each gene cluster derived in the data preprocessing step. In step 3) we use Hamming distances as a dissimilarity measure between partitions. Partitions $\mathcal{P}_i$ are viewed as N-vectors $\boldsymbol{p}_i$ containing the binary cluster labels taking values from {0,1}. In order to make Hamming distances suitable for this purpose, we have to consider the inherent permutation symmetry of the clustering process: a cluster called "1" in the first partition can be called "0" in the second one. When computing the pairwise Hamming distances, we thus have to minimize over the two possible permutations π of cluster labels

$$d_{\mathrm{Hamming}}(\mathcal{P}_i, \mathcal{P}_j) = \min_{\pi} \sum_{k=1}^{N} \left| \boldsymbol{p}_i(k) - \pi\left(\boldsymbol{p}_j(k)\right) \right|. \qquad (6)$$

Steps (iv) and (v) need some further explanation: the problem of identifying compact groups in datasets which are represented by pairwise distances can be solved by optimizing the *pairwise clustering cost function*, [20]. We iteratively increase the number of clusters (which is a free parameter in the pairwise clustering functional) until the average dissimilarity in each group does not exceed a predefined threshold. Reasonable problem-specific thresholds can be defined by considering the following null-model: given N samples, the average Hamming distance between two randomly drawn 2-partitions $\mathcal{P}_1$ and $\mathcal{P}_2$ is roughly $d_{\mathrm{Hamming}}(\mathcal{P}_1, \mathcal{P}_2) \approx N/2$. It may thus be reasonable to consider only clusters which are several times more homogeneous than the expected null-model homogeneity.

For the clusters which are considered homogeneous, we observe their populations, and out of all models investigated we choose the one leading to the partition cluster of largest size. For this dominating cluster, we then select a prototypical partition. For selecting such prototypical partitions in pairwise clustering problems, we refer the reader to [21], where it is shown that the pairwise clustering problem can be equivalently restated as a k-means problem in a suitably chosen embedding space. Each partition is represented as a vector in this space. This property allows us to select those partitions as representants, which are closest to the partition cluster centroids. The whole work-flow of model selection is summarized schematically in Fig. 1.

IV. A Demo Application: The AML/ALL Dataset

The Leukemia data set published by Golub *et al.*, [2], consists of 72 samples, of which 47 are *acute lymphoblastic leukemia* (ALL), and 25 are *acute myeloid leukemia* (AML). Expression levels of 7129 genes were measured using Affymetrix arrays.

This dataset has been analyzed before in various papers, and we are aware that another study may be of limited interest from a biological perspective. Nevertheless, we have decided to include this experiment here, since in our opinion a novel method—and

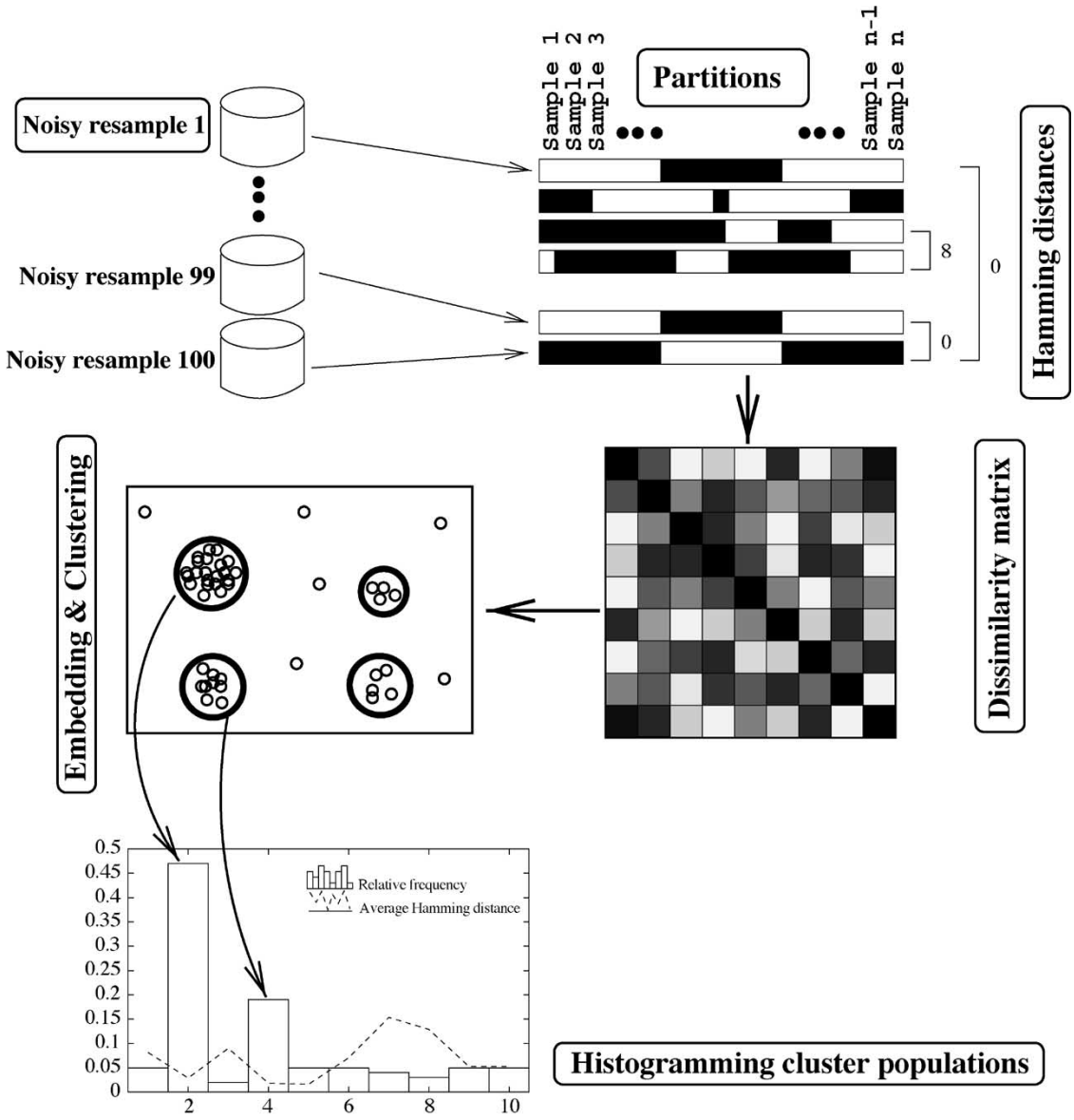

Fig. 1. Model selection by resampling: schematic overview of the selection work-flow for one fixed value of the ℓ_1-constraint κ. This process is repeated for several values of κ, and the one which leads to the highest populated partition cluster is finally selected.

in particular a novel unsupervised clustering method—should be first validated in a control experiment. Moreover, this dataset is highly suited for demonstrating the whole process flow of our class discovery method.

A. Data Preprocessing

In a first standard preprocessing step, gene expression values were subjected to a variation filter that excluded genes showing minimal variation across the samples being analyzed. We excluded genes with $max/min < 2$ and $max - min < 1000$, leading to a reduced set of 1479 genes. Then, the data were log-transformed, centered to zero mean, "squashed" through a $tanh$-function for outlier-reduction, and standardized to unit variance (for each microarray).

Following [22], in a next step we extracted the 200 genes with highest variance across the samples. While the number 200 might appear completely artificial, in [22] it has been shown that for both datasets we have analyzed in this paper, this number roughly corresponds to a "knee" in the variance plot which separates high-variance genes from a flat "bulk" spectrum. Instead of directly using these 200 genes, however, we decided to first cluster the genes into 40 compact groups by using the k-means algorithm. The reasons for this preclustering are the following: 1) working with the cluster prototypes instead of with the original genes has the potential to average out the noise in the dataset (note that multiplicative intensity-dependent noise has been log-

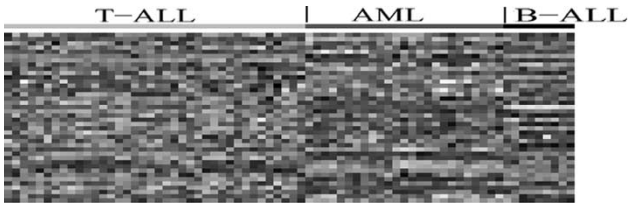

Fig. 2. Expression patterns of the 40 cluster representants. Bright grey-values represent high expression levels.

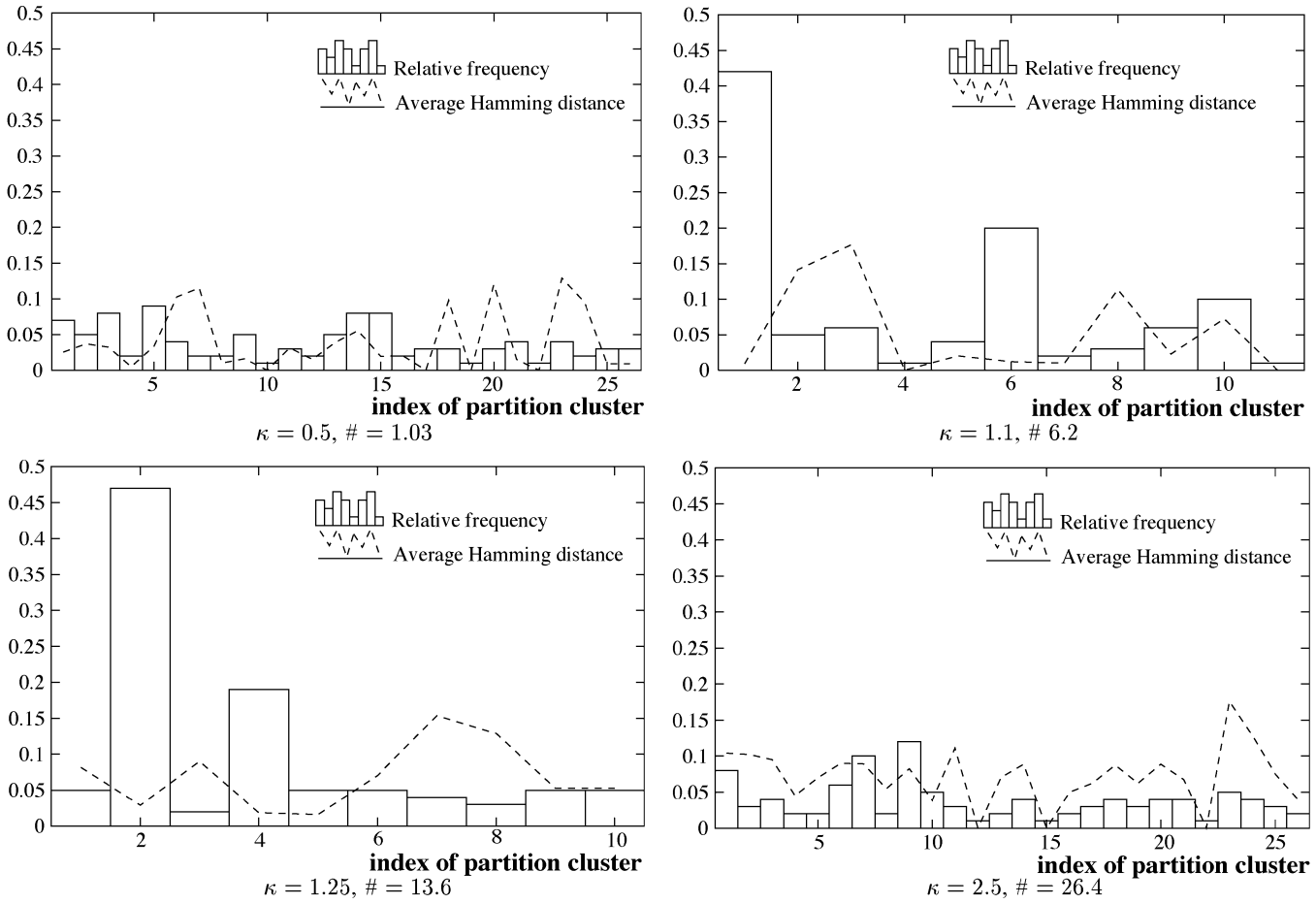

Fig. 3. Clustering of the 100 re-sampled datasets. The histograms depict the relative frequency of the individual groups, which are consecutively numbered on the horizontal axes. The dashed curves indicate the average Hamming distance in each cluster (multiplied by 0.01). The Hamming distance is plotted in the form of a curve for visualization purposes only: there is no intentional ordering of the cluster indices on the horizontal axis. κ denotes the constraint value, and # the average number of selected features.

transformed to additive noise); 2) clustering avoids collinearities in the data which are problematic for the covariance estimation; 3) it speeds up the computations without loosing too much information: the gene clusters obtained are very homogenous, and variations within the clusters may be readily explained by noisy measurements. The homogeneity of the gene clusters can be quantified by considering a null-model: the expression levels on each chip are standardized to have unit variance. Thus, for a random partitioning of the genes, we would expect to observe the same variance of one within each cluster. The observed variance within the 40 gene clusters, however, turned out to be only 0.2, and thus five times lower than the expected variance under the random model. While the choice of 40 clusters still appears somewhat artificial, our experiments at least show that on a broad range between 30 and 60 clusters the results are highly similar.

Since in our model genes within one gene-cluster are considered indistinguishable, the value of the within-cluster variance defines the level of noise by which each expression measurement is corrupted. Moreover, this variance also defines a suitable noise level for *artificially* corrupting the dataset when drawing noisy resamples for selecting the optimal ℓ_1-constraint value according to the model selection procedure described in Section III.

The finally chosen expression patterns of the 40 cluster representants (the centroid-vectors) are depicted in Fig. 2. In the remainder of this paper, these patterns will be simply called "features."

The goal of our class discovery algorithm can now be stated as simultaneously finding sample partitions and automatically extracting a subset of features which are most discriminative for these partitions.

B. Model Selection and Experimental Evaluation

Fig. 3 depicts the outcomes of the resampling-based stability analysis described in Section III. For different constraint—values on the interval [0.5, 2.5], we draw 100 noisy data resamples, run our class discovery algorithm, compute the pairwise Hamming distances between partitions, and group the partitions into homogeneous clusters.

The AML/ALL dataset contains 72 samples, so that the expected Hamming distance between two *randomly* drawn 2-partitions is roughly $d^{\text{rand}} \approx N/2 = 36$. In our experiments we considered only those partition-cluster as "homogeneous," which have an average Hamming distance $\bar{d} < 1/2 \cdot d^{\text{rand}} = 18$. We found, however, that the selection mechanism is rather insensitive to value: the most populated clusters turned out to be very

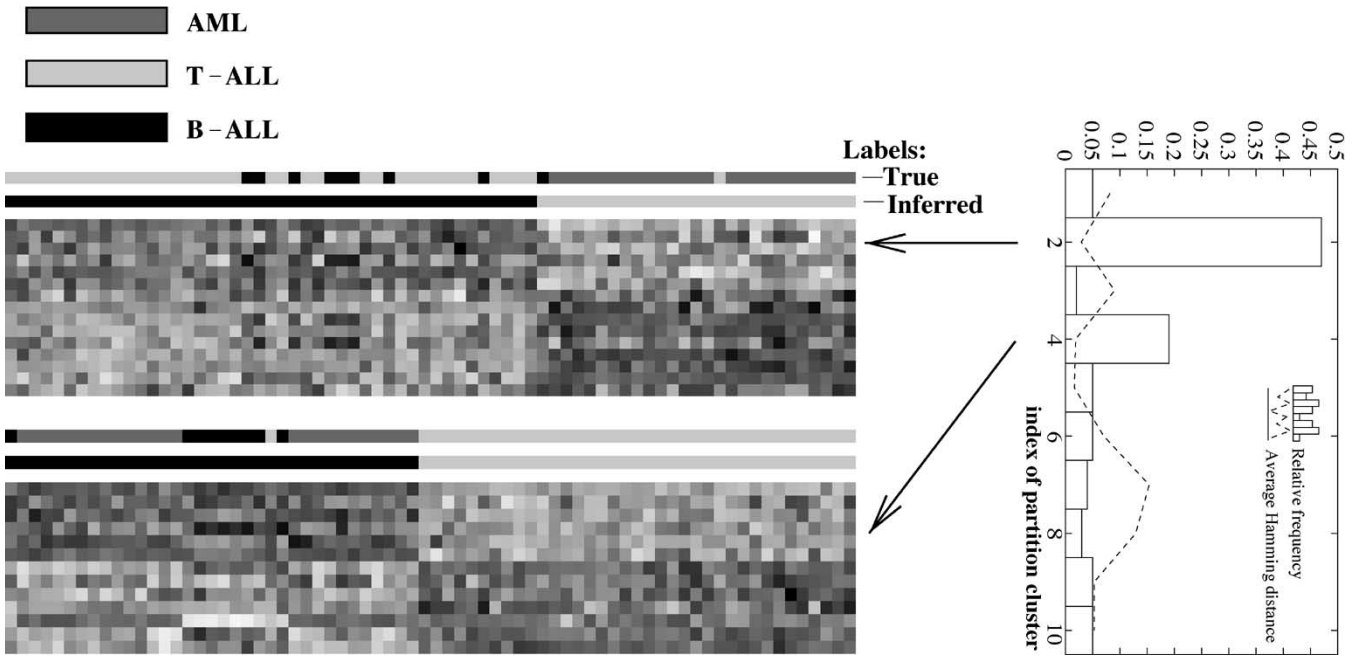

Fig. 4. Representants of the two dominant partitions for $\kappa = 1.25$. Samples are ordered w.r.t. the inferred labels (lower label indicator bar). Upper bar: original "true" labels. The automatically selected features have been ordered w.r.t. over-expression for one of the two classes.

homogeneous ($\bar{d} < 4$, see upper right panel), so that a decrease in the homogeneity constraint only leads to a further splitting of the clusters with the lowest population. The latter, however, do not affect the parameter selection at all.

For $\kappa = 0.5$ (upper left panel) we need 26 clusters to satisfy our *a priori* chosen constraint about the maximum inhomogeneity in each cluster. According to the histogram, no cluster contains more than 10% of the partitions. There is no distinct correlation between the average Hamming distance in each cluster and frequency. For this value of κ, in each partition on average 1.03 features have been selected (recall that our features are representatives of gene clusters). A very similar situation occurs for $\kappa = 2.5$ (lower right panel), where we have on average 26.4 features (i.e. more than halve of all features have been selected). Between these two extremes, however, we observe stable partitions. For $\kappa = 1.25$ (lower left panel) we observe two highly stable partitions: 47% of the partitions belong to the group labeled "2," and 19% to group "4." Note that in this case we only need 10 cluster to model all 100 partitions, while satisfying the homogeneity constraint $\bar{d} < 18$. In each partition on average 13.6 features have been selected. Moreover, there is a distinct correlation of high homogeneity and high frequency, indicating that the clusters contain many highly similar partitions. The plot for $\kappa = 1.1$ (upper right panel) shows that over a relatively broad range of κ-values the partition diagram varies smoothly: the prototypical partition for partition cluster "1" in the upper right panel is identical to the respective prototype for cluster "2" in the lower left panel. The same identity between partitions is observed for the clusters with the second highest population. This result indicates that even a relatively coarse grid-search procedure for selecting κ should be capable of detecting stable partitions.

Fig. 4 indicates that these two stable partitions have a clear biological meaning: the most dominant split separates AML samples from ALL samples (with two errors). The second split separates T-cell ALL from both B-cell ALL and AML (one error).

Having identified a constraint value which leads to stable partitions, we now turn our attention to the **stability of the feature selection process**: for the 47 partitions belonging to the most dominant split we count how often each of the features has been automatically selected. The feature-selection mechanism turns out to be very stable, too: 12 features are selected with a frequency higher than 0.5, five of which with a frequency higher than 0.8. These five most frequently selected features (i.e. the members of the corresponding gene clusters) are shown in Table I.

Concerning the separation of the two classes of Leukemia, it is interesting to notice that the discriminative power of most of the high-scoring genes in Table I may have a biological interpretation: In the **first cluster**, we find Cyclin D3, which has been identified as a dominant oncogene in the pathogenesis and transformation in several histologic subtypes of mature B-cell malignancies, [23]. IL7R is a member of the interleukin receptor family. Expression of interleukin receptors (in particular expression of IL2R, which appears in our list in the 12th cluster) has been examined on a wide range of cells of myeloid origin including bone marrow blasts obtained from acute myelogenous leukemia (AML) patients, [24].

It is also interesting to find the interferon stimulated gene HEM45 (ISG20) among the top-scoring genes. ISG20 is one of the nuclear bodies (NBs)-associated proteins, which could play an important role in oncogenesis and viral infections, [25]. According to [26], expression of the probable G protein-coupled receptor LCR1 homolog (alias CD184 antigen) is associated with survival in familial chronic lymphocytic leukemia.

The **second cluster** contains e.g. the proliferation-associated gene PAGA (alias natural killer-enhancing factor A). It is involved in redox regulation of the cell and might participate in the signaling cascades of growth factors and tumor necrosis factor-alpha. In the **third cluster** we find the CD37 antigene, which has been shown to provide highly significant discrimination between chronic lymphocytic leukemia (CLL) and normal periph-

TABLE I
AML/ALL DATASET: HIGH-SCORING GENE CLUSTERS FOR THE MOST DOMINANT SPLIT. FIRST COLUMN: NUMBER OF FEATURE (GENE-CLUSTER). SECOND COLUMN: FREQUENCY SCORE OF FEATURE

#	Score	Cluster members
1	0.94	AIM, Macmarcks, IL7R, CCND3, HEM45, LCR1
2	0.91	SLIM1, DPYSL2, ALDR1, LMP2, PAGA
3	0.85	CD37, BSG, NM23D, P4HB, EIF5A
4	0.83	CD33, BB1, TGFB1, GliPR
5	0.81	HSPA8, PSMA6, c-MYB, TCF3(E2A), hSNF2b

eral blood leukocytes, [27]. A member of the **fourth cluster** is the differentiation antigene CD33. According to [28] it possesses high expression specificity to AML. On the contrary, TCF3 (alias E2A) in the **fifth cluster** has known expression specificity to ALL, [29]. Moreover, the fifth cluster also contains the well-known proto-onkogene c-MYB. Among the **clusters no. 6–12**, we also find the well-known marker antigenes CD19 and CD63. The latter is a widely expressed glycoprotein member of the TM4SF superfamily that is present on many non-lymphoid cells, [30].

C. Comparison With Other Algorithms

Some other class discovery algorithms have been tested on the AML/ALL Leukemia dataset. BEN-DOR *et al.*, [4], report on the splitting of the dataset into two subsets according to different information theoretic scoring measures, which are used as "building blocks" in their algorithm. It is interesting to note that for all scores used, the three highest scoring labelings are significantly different from the true AML/ALL labels, and only the fourth inferred labeling is similar to the true one. Compared with our method, their results are highly ambiguous in the sense that the user is left with several possible high-scoring solutions, from which he/she can only pick one by chance. Even worse, the highest-scoring labelings do not reveal the true structure of the samples.

In [5], [6] a *support vector machine*-based class discovery algorithm has been presented and tested on this dataset. Among the ten highest scoring partitions in [6], the 4th and the 10th partition separate the dataset into two groups which are similar to the true AML/ALL labels. The 1st partition separates B-cell ALL from the two other classes. Although their method is capable of finding splits which are similar to the ground-truth labeling, the SVM-based algorithm also suffers from ambiguities: among the ten highest scoring partitions, there are six which seem to *impose artificial structure* rather than reconstructing the original labels.

We conclude, that compared to other approaches reported in the literature, the method presented in this paper combines two outstanding features: the class discovery algorithm is capable of reproducing the true structure hidden in the dataset, and our stability-based model selection strategy successfully overcomes the problem of ambiguous solutions.

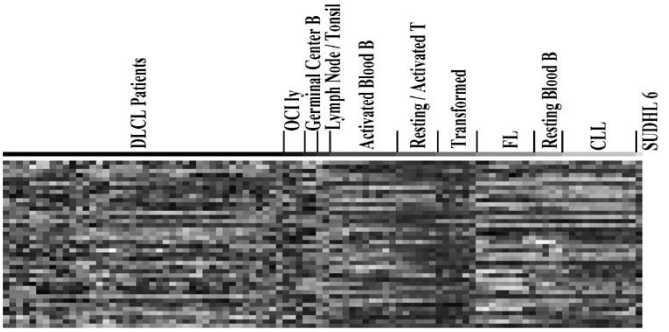

Fig. 5. DLBCL dataset: expression profiles of the 40 gene-cluster representants. The samples are ordered with respect to their membership in either the group of *diffuse large B-cell lymphoma* (DLBCL) patients or in one of several control groups from different cell lines.

V. CLASS DISCOVERY IN LARGE B-CELL LYMPHOMA SAMPLES

In a second experiment we re-analyzed the dataset published by ALIZADEH *et al.*, [1]. We used the same data-preprocessing as in the preceding example, i.e. we extracted the 200 genes with highest variance across the samples and preclustered these genes into 40 gene-clusters (our features from which we want to select indicative subsets). The representants of these gene-clusters are depicted in Fig. 5.

A. The First Hierarchical Split

Fig. 6 shows the first split of the whole dataset into two subgroups. The left panel presents the population of the partition clusters for the optimal value of the ℓ_1-constraint $\kappa = 0.9$, which has been chosen via a grid search on the interval [0.5,1.5]. The reader should notice, that the stability plot suggests that the dominating cluster no. 2 is highly homogenous, and that it defines the only stable separation of the samples into two groups. On average 3.06 features (gene-clusters) have been automatically selected. The corresponding prototypical partition is depicted in the right panel. According to the color-coded indicator bar of ground-truth labels, this split is highly correlated to the separation of *diffuse large B-cell lymphoma* (DLBCL) patients from all the other samples. There are only four exceptions: the cluster labeled *DLBCL* contains the two Lymph-node/Tonsil samples, and the *Non-DLBCL* cluster contains the samples DLCL0042 and DLCL0009.

Most members of the automatically selected gene clusters in Table II belong to a group of genes which has been classified in [1] as defining a "lymph-node" signature. This class of genes includes genes with high expression specificity to macrophages, like CSF-1 and the small inducible cytokine A5 (RANTES). Also in this category falls the allograft inflammatory factor-1 (IBA1), which is a bioactive macrophage factor which might play a role in macrophage activation and function, [31].

Also in accordance with the analysis in [1], we find several genes involved in remodeling the extracellular matrix, such as the matrix metalloproteinases MMP-9 and MMP-2, the metalloproteinase inhibitor TIMP-3, and SPARC. The latter appears to regulate cell growth and is capable of delaying tumor growth *in vivo*, [32].

The list of high-scoring genes also includes several genes which are known to be associated with tumor progression/invasion, like the CD63 antigene and human cathepsin B. The latter

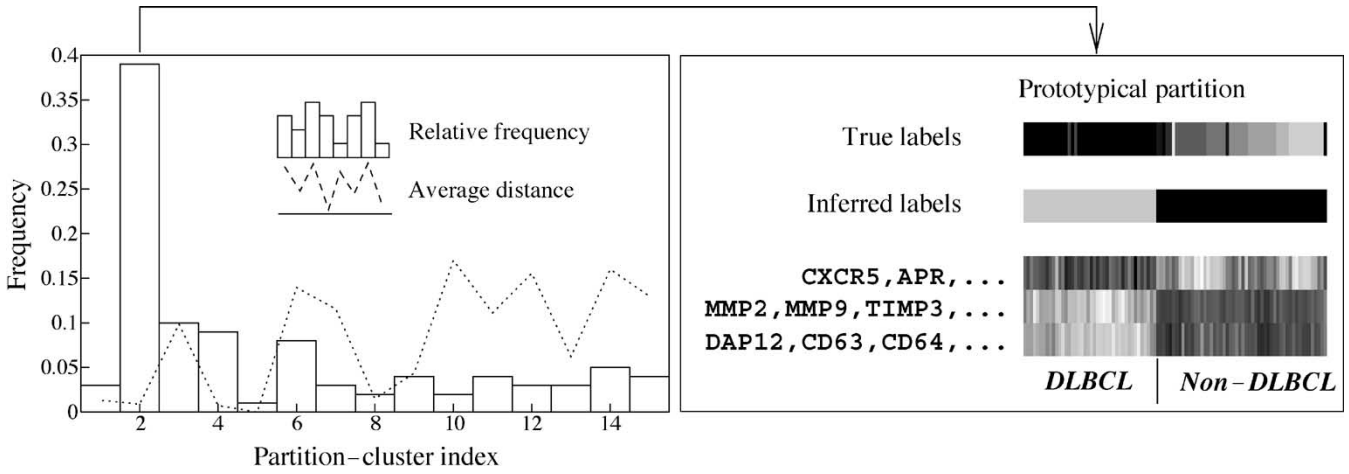

Fig. 6. Left panel: population of partition-clusters for the optimal value of the ℓ_1-constraint $\kappa = 0.9$. Right panel: prototypical partition for the dominating partition cluster no. 2 and automatically selected gene-clusters.

TABLE II
AML/ALL DATASET: HIGH-SCORING GENE CLUSTERS AND ANNOTATED GENES WITHIN THE CLUSTERS FOR THE MOST DOMINANT SPLIT OF THE DATA SET. FIRST COLUMN: FREQUENCY SCORE OF GENE CLUSTER

1.0	DAP12, CD63, CD64, FCRI, FCERI, RANTES, cathepsin B, cathepsin L, IBA 1, Cyclin D2
0.97	CXCR5, APR
0.97	MMP-2, MMP-9, CSF-1, Cytochrome P450, RUNX2(OSF-2), SPARC, Fibronectin 1, TIMP-3
0.26	SLC, IP-10, Humig, Guanylate binding protein 1

is a proteolytic enzyme implicated in tumor invasion and metastasis, [33]. We also find CXCR5 which is one component of the chemokine/chemokine receptor pair CXCL13/CXCR5 that is required for the architectural organization of B cells within lymphoid follicles, [34].

Moreover, the first gene-cluster in Table II contains cyclin D2, a proto oncogene belonging to the class of D-type cyclins. These genes are involved in key cellular decisions that control cell proliferation, cell-cycle arrest, quiescence, and differentiation, [35].

B. Refining the Partition: Discovery of DLBCL Subtypes

Having found a stable partition of the samples into a DLBCL cluster and a non-DLBCL cluster, we now investigate further refinements of this partition.

In the original analysis of this dataset in [1], several distinct expression *signatures* have been identified. The authors mentioned that "... in principle each of these gene expression signatures could be used to define subsets of DLBCL. We decided to focus our attention initially on the germinal centre B-cell genes ...." This statement of the authors addresses the inherent problem of ambiguities in class discovery, which in their analysis has been overcome by exploiting biomedical *a priori* knowledge. The restriction to the subset of genes showing a *germinal centre B-cell signature* led to the discovery of two subgroups of DLBCL, which turned out to add to the prognostic value of a standard clinical indicator of prognosis. In the following we describe a purely statistical approach to this problem, that does not depend on any kind of prior knowledge. Our automatically found subgroups are shown to add to the prognostic value in the same sense and at the same confidence level as the originally defined subgroups.

In Fig. 7 the most stable split of the *DLBCL* cluster is depicted. The optimal ℓ_1-constraint value was $\kappa = 1.2$. On average 12.2 features (gene clusters) have been automatically selected by the Bayesian relevance determination mechanism. Details of the consensus split induced by the prototypical partition are depicted in the lower panel. We decided to name the first group of samples the *LN*-cluster, since it contains the two Lymph-Node/Tonsil samples, and the second group the *Non-LN* cluster, respectively. The overall correspondence of these two clusters with the DLBCL subgroups found in [1]—*GC B-like* and *activated B-like*—is about 73%. This agreement may be viewed more as a tendency than as a significant correlation (for a random partition we would expect 50% agreement). Despite the differences in splitting the DLBCL samples, it turns out, however, that our *LN/Non-LN* partition adds to the prognostic value of a standard clinical prognostic factor in the same sense as the *GC B-like/activated B-like* partition. Before going into details on the prognostic issue, we now briefly discuss the genes that have been automatically selected (Table III).

The occurrence of most of these genes has a clear biological interpretation. A primal source of lymphomas is disruption of the regulation of B-cell differentiation and activation, resulting in oncogenic chromosomal translocations that block differentiation, prevent apoptosis and/or promote proliferation. In DLBCLs, for instance, apoptosis can be abrogated by translocation, amplification or transcriptional activation of the BCL-2 gene, [36]. It is interesting to observe expression of BCL-2 in many of the *LN*-samples, see Fig. 7. Patients in this group with low IPI score had a distinctly worse overall survival than patients

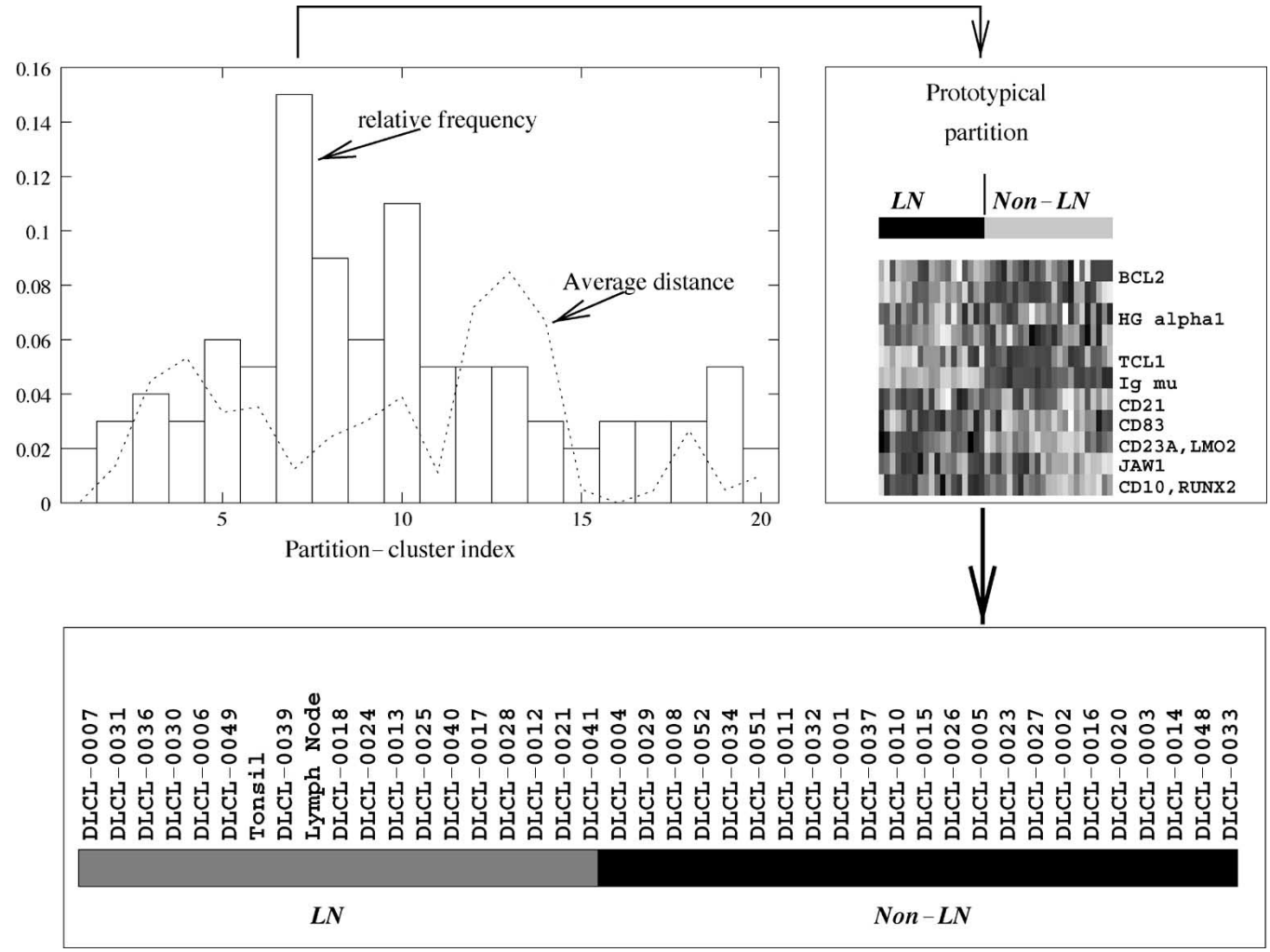

Fig. 7. Splitting the DLBCL cluster into one group named *LN* which contains the Lymph-Node/Tonsil samples, and into a *Non-LN* group.

TABLE III
DLBCL SUBGROUPS: HIGH-SCORING GENE CLUSTERS AND ANNOTATED GENES WITHIN THESE CLUSTERS

1.0	TCL-1
1.0	Immunoglobulin mu
0.67	CD23A, SA3, SLAM, LMO2(TTG-2)
0.67	JAW1
0.67	FLAP, CD10, RUNX2(OSF-2)
0.53	BCL-2

in the *Non-LN* group, cf. [8]. This observation may be corraborated by HAYASHI *et al.*, [37], who report a tendency, in which patients with BCL-2 overexpression resulted in poor prognosis in the case of primary central nervous system lymphomas (PCNSLs) of the diffuse large B-cell type. Among the high-scoring genes, we also find the B-cell specific antigene CD23a, which has a essential role in the differentiation of B-cells.

The highest scoring gene is TCL1 which is overexpressed in the *LN* group. According to [38], the TCL1 protooncogene is overexpressed in many mature B cell lymphomas, especially from AIDS patients. For Non-AIDS-related lymphomas, it has been reported in [39] that TCL1 expression in B cell lymphoma usually reflects the stage of B cell development from which they derive.

In accordance with the grouping proposed in [1], we also find some Germinal-centre B-cell signature genes among the most discriminative genes—for example CD10 and JAW1.

C. DLBCL Subgroups and Prognostic Categories

The left panel of Fig. 8 presents Kaplan-Meier plots of overall survival data from the DLBCL patients, segregated according to low and high values of the International Prognostic Indicator (IPI). This clinical indicator of prognosis takes into account the patient's age, performance status, and the extent and location of disease, cf. [40]. As can be seen in the Kaplan-Meier plot, the two IPI classes (low and high risk) are associated with statistically significant differences in overall survival ($P = 0.03$ in a log-rank test, see [41]). Following [1], we tested the specifity of our inferred partitioning to overall survival within the IPI *low risk* group (IPI score 0–2). The right panel of Fig. 8 presents the corresponding Kaplan-Meier plot for these low risk DLBCL patients, segregated according to cluster membership in either the *LN* or the *Non-LN* cluster. The *LN* cluster turns out to be associated with relatively low survival probability, whereas the

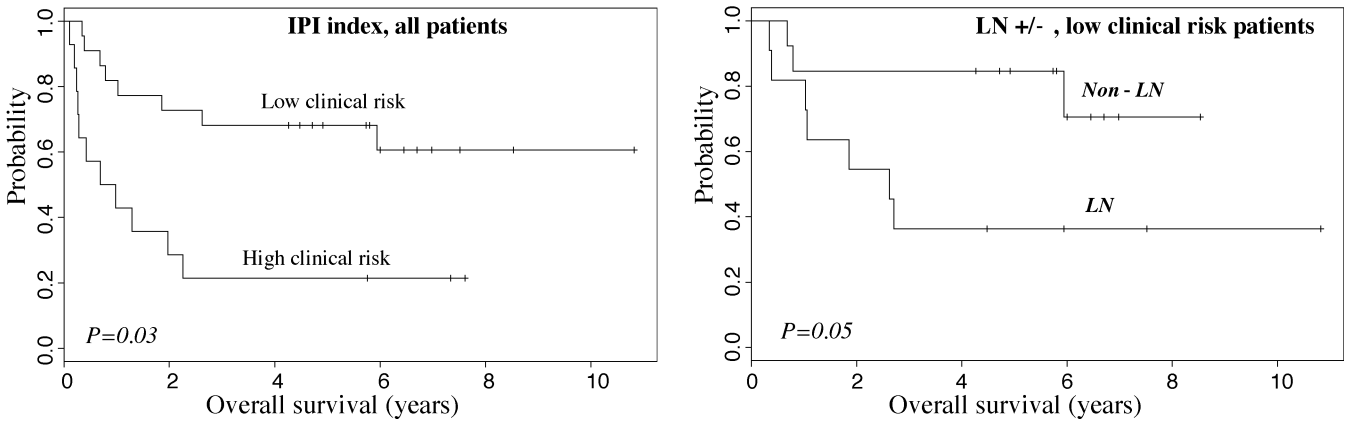

Fig. 8. DLBCL subgroups. Left panel: Kaplan-Meier plot of overall survival of DLBCL patients grouped according to the International Prognostic Index (IPI). Low clinical risk patients (IPI score 0–2) and high clinical risk patients (IPI score 3–5) are plotted separately. Right panel: Kaplan-Meier plot of overall survival of low clinical risk DLBCL patients (IPI score 0–2) grouped on the basis of the automatically inferred clusters *LN* and *Non-LN*.

Non-LN cluster defines a subgroup of DLBCL patients with very high survival probability. According to a log-rank test, the differences in overall survival times between the two subgroups are significant at a $P = 0.05$ level.

A comparison with the experimental results in the original paper by ALIZADEH *et al.* suggests that our automatically inferred subgroups add to the prognostic value of the IPI indicator in the same sense and at the same confidence level as the *GC B-like* and *activated B-like* subgroups proposed in [1]. The reader should notice, however, that the latter subgroups have been identified after *manually* selecting a certain subset of genes showing a *germinal centre B-cell* signature. Our approach, on the contrary, is a pure statistical approach. Therefore, it has the potential of dealing with a larger class of problems for which prior knowledge about gene function might not be available.

VI. Conclusion

The problem of *class discovery* in microarray experiments consists of simultaneously finding distinct groups of samples and automatically extracting subsets of features which are most discriminative for these partitions. Some approaches to this problem have been proposed in the literature, most of which, however, bear several inherent shortcomings, such as an unclear probabilistic model, the simplifying assumption of features as being uncorrelated, or the absence of a plausible model selection strategy. The latter issue is of particular importance, since many approaches suffer from ambiguities caused by contradictory splitting hypotheses. In this work we have presented a new approach to class discovery which has the potential to overcome these shortcomings. It has a clear interpretation in terms of a constrained Gaussian mixture model, which combines a clustering method with a Bayesian inference mechanism for automatically selecting relevant features. The relevance determination mechanism has been incorporated in the M-step of the classical EM-algorithm for Gaussian mixtures. Thus, both basic ingredients of our class discovery method, namely *clustering* and *feature selection*, optimize the same objective function.

We further present an optimization algorithm with guaranteed convergence to a local optimum. This optimization algorithm has only one free parameter (the value of the ℓ_1-constraint), for which we propose a stability-based model selection procedure: by drawing noisy re-samples from the dataset, we identify models which lead to partitions that are stable both with respect to noise in the data and with respect to numerical optimization problems caused by multiple local optima. For each model considered, we simultaneously analyze the stability of the feature selection process involved. Experiments with real-world datasets effectively demonstrate that this class discovery method is able to correctly infer partitions and corresponding genes which are both relevant in a biological sense. For the task of inferring subgroups of B-cell Lymphoma patients, we have shown that the partitions found by our class discovery method add to the prognostic value of a standard clinical prognosis indicator.

In order to make this method available to others, we are currently working on a user-friendly and stable software solution. For a copy of our "prototypical" software with which the results in this paper have been produced, please contact the authors.

Acknowledgment

The authors would like to thank J. M. Buhmann and M. Braun for substantial and helpful discussions.

References

[1] A. Alizadeh, M. Eisen, R. Davis, C. Ma, I. Lossos, A. Rosenwald, J. Boldrick, H. Sabet, T. Tran, X. Yu, J. Powell, L. Yang, G. Marti, T. Moore, J. Hudson, W. Chan, T. Greiner, D. Weissenberger, J. Armitage, R. Levy, M. Grever, J. Byrd, D. Botstein, P. Brown, and L. Staudt, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, pp. 503–511, 2000.

[2] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531–537, Okt. 15, 1999.

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

[4] A. Ben-Dor, N. Friedman, and Z. Yakhini, "Class discovery in gene expression data," in *Proc. RECOMB*, 2001, pp. 31–38.

[5] A. v.Heydebreck, W. Huber, A. Poustka, and M. Vingron, "Identifying splits with clear separation: a new class discovery method for gene expression data," *Bioinformatics*, vol. 17, pp. 107–114, 2001.

[6] F. Markowetz and A.V. Heydebreck, "Class discovery in gene expression data: characterizing splits by support vector machines," in *Proc. 26th Ann. Conf. Gesellschaft für Klassifikation*, 2002.
[7] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *JRSS B*, vol. 39, pp. 1–38, 1977.
[8] T. Hastie and R. Tibshirani, "Discriminant analysis by Gaussian mixtures," *JRSS B*, vol. 58, pp. 158–176, 1996.
[9] T. Hastie, R. Tibshirani, and A. Buja, "Flexible discriminant analysis by optimal scoring," *JASA*, vol. 89, pp. 1255–1270, 1994.
[10] T. Hastie, A. Buja, and R. Tibshirani, "Penalized discriminant analysis," *AnlsStat.*, vol. 23, pp. 73–102, 1995.
[11] R. Tibshirani, "Regression shrinkage and selection via the lasso," *JRSS B*, vol. 58, no. 1, pp. 267–288, 1996.
[12] V. Roth, "The Generalized Lasso: A Wrapper Approach to Gene Selection for Microarray Data," Univ. Bonn, Dept. Computer Science III, Tech. Rep. IAI-TR-2002-8, 2002.
[13] ——, "The generalized LASSO," *IEEE Trans Neural Networks*, vol. 15, Jan. 2004.
[14] D. MacKay, "Bayesian nonlinear modeling for the prediction competition," in *ASHRAE Trans. Pt. 2*, vol. 100, Atlanta, GA, 1994, pp. 1053–1062.
[15] M. Figueiredo and A. K. Jain, "Bayesian learning of sparse classifiers," in *Proc. CVPR200I*, 2001, pp. 35–41.
[16] M. Osborne, B. Presnell, and B. Turlach, "On the lasso and its dual," *JCmpGrSt*, vol. 9, pp. 319–337, 2000.
[17] F. Meinecke, A. Ziehe, M. Kawanabe, and K.-R. Müller, "A resampling approach to estimate the stability of one- or multidimensional independent components," *IEEE Trans. Biomed. Eng.*, vol. 49, pp. 1514–1525, Dec. 2002.
[18] S. Harmeling, F. Meinecke, and K.-R. Müller, "Injecting noise for analysing the stability of ICA components," *Signal Processing*, vol. 84, pp. 255–266, 2004.
[19] T. Lange, M. Braun, V. Roth, and J. Buhmann, "Stability-based model selection," in *Advances in Neural Information Processing Systems*. Cambridge, MA: M.I.T. Press, 2002, vol. 15, pp. 617–624.
[20] T. Hofmann and J. Buhmann, "Pairwise data clustering by deterministic annealing," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 1–14, Jan. 1997.
[21] V. Roth, J. Laub, J. M. Buhmann, and K.-R. Müller, "Going metric: denoising pairwise data," in *Advances in Neural Information Processing Systems*. Cambridge, MA: M.I.T. Press, 2003, vol. 15, pp. 817–824.
[22] S. Dudoit and J. Fridlyand, "A prediction-based resampling method for estimating the number of clusters in a dataset," *Genome Biol.*, vol. 3, no. 7, 2002.
[23] T. Sonoki, L. Harder, D. Horsman, L. Karran, I. Taniguchi, T. Willis, S. Gesk, D. Steinemann, E. Zucca, B. Schlegelberger, F. Sole, A. Mungall, R. Gascoyne, R. Siebert, and M. Dyer, "Cyclin D3 is a target gene of t(6;14)(p21.1;q32.3) of mature B-cell malignancies," *Blood*, vol. 98, no. 9, pp. 2837–2844, 2001.
[24] R. Schumann, T. Nakarai, H. Gruss, M. Brach, U. von Arnim, C. Kirschning, L. Karawajew, W. Ludwig, J. Renauld, J. Ritz, and F. Herrmann, "Transcript synthesis and surface expression of the interleukin-2 receptor (alpha-, beta-, and gamma-chain) by normal and malignant myeloid cells," *Blood*, vol. 87, no. 6, pp. 2419–2427, 1996.
[25] M. Chelbi-Alix and H. de The, "Herpes virus induced proteasome-dependent degradation of the nuclear bodies-associated PML and Sp100 proteins," *Oncogene*, vol. 18, no. 4, pp. 935–941, 1999.
[26] N. Ishibe, M. Albitar, I. Jilani, L. Goldin, G. Marti, and N. Caporaso, "CXCR4 expression is associated with survival in familial chronic lymphocytic leukemia, but CD38 expression is not," *Blood*, vol. 100, no. 3, pp. 1100–1101, 2002.
[27] L. Belov, O. de la Vega, C. dos Remedios, S. Mulligan, and R. Christopherson, "Immunophenotyping of leukemias using a cluster of differentiation antibody microarray," *Cancer Res.*, vol. 61, no. 11, pp. 4483–4489, 2001.
[28] R. Sobol, R. Mick, I. Royston, F. Davey, R. Ellison, R. Newman, J. Cuttner, J. Griffin, H. Collins, and D. Nelson, "Clinical importance of myeloid antigen expression in adult acute lymphoblastic leukemia," *N. Engl. J. Med.*, vol. 316, pp. 1111–1117, 1987.
[29] J. Nourse, N. Galili, J. Wilkinson, E. Stanbridge, S. Smith, and M. Cleary, "Chromosomal translocation t(1:l9) results in synthesis of a homeobox fusion mRNA that codes for a potential chimeric transcription factor," *Cell*, vol. 60, no. 4, pp. 535–545, 1990.
[30] D. Smith, P. Monk, and L. Partridge, "Antibodies against human CD63 activate transfected rat basophilic leukemia (RBL-2H3) cells," *Mol. Immunol.*, vol. 32, no. 17–18, pp. 1339–1344, 1995.
[31] K. Watano, K. Iwabuchi, S. Fujii, N. Ishimori, S. Mitsuhashi, M. Ato, A. Kitabatake, and K. Onoe, "Allograft inflammatory factor-1 augments production of interleukin-6, -10 and -12 by a mouse macrophage line," *Immunology*, vol. 104, no. 3, pp. 307–316, 2001.
[32] C. Schultz, N. Lemke, S. Ge, W. Golembieski, and S. Rempel, "Secreted protein acidic and rich in cysteine promotes glioma invasion and delays tumor growth *in vivo*," *Cancer Res.*, vol. 62, no. 21, pp. 6270–6277, 2002.
[33] J. MacKenzie, S. Mason, J. Hickford, M. Kohonen-Corish, and R. Bickerstaffe, "A polymorphic marker for the human cathepsin B gene," *Mol. Cell Probes*, vol. 15, no. 4, pp. 235–237, 2001.
[34] H. Husson, A. Freedman, A. Cardoso, J. Schultze, O. Munoz, G. Strola, J. Kutok, E. Carideo, R. De Beaumont, F. Caligaris-Cappio, and P. Ghia, "CXCL13 (BCA-1) is produced by follicular lymphoma cells: role in the accumulation of malignant B cells," *Br. J. Haematol.*, vol. 119, no. 2, pp. 492–495, 2002.
[35] J. Bartkova, E. Rajpert-de Meyts, N. Skakkebaek, and J. Bartek, "D-type cyclins in adult human testis and testicular cancer: relation to cell type, proliferation, differentiation, and malignancy," *J. Pathol.*, vol. 187, no. 5, pp. 573–581, 1999.
[36] A. Shaffer, A. Rosenwald, and L. Staudt, "Lymphoid malignancies: the dark side of B-cell differentiation," *Nature Rev. Immunol.*, pp. 920–933, 2002.
[37] Y. Hayashi, M. Iwato, Y. Arakawa, H. Fujisawa, Y. Thoma, M. Hasegawa, O. Tachibana, and J. Yamashita, "Homozygous deletion of INK4aIARF genes and overexpression of bcl-2 in relation with poor prognosis in immunocompetent patients with primary central nervous system lymphoma of the diffuse large B-cell type," *J. Neurooncol.*, vol. 55, no. 1, pp. 51–58, 2001.
[38] K. Hoyer, S. French, D. Turner, M. Nguyen, M. Renard, C. Malone, S. Knoetig, C. Qi, T. Su, H. Cheroutre, R. Wall, D. Rawlings, H. Morse, and M. Teitell, "Dysregulated TCLI promotes multiple classes of mature B cell lymphoma," *PNAS*, vol. 99, no. 22, pp. 14 392–14 697, 2002.
[39] J. Said, K. Hoyer, S. French, L. Rosenfelt, M. Garcia-Lloret, P. Koh, T. Cheng, G. Sulur, G. Pinkus, W. Kuehl, D. Rawlings, R. Wall, and M. Teitell, "TCLI oncogene expression in B cell subsets from lymphoid hyperplasia and distinct classes of B cell lymphoma," *Lab. Investigat.*, vol. 81, no. 4, pp. 555–564, 2001.
[40] M. Shipp *et al.*, "The international nonhodgkin's lymphoma prognostic factors project: a predictive model for aggressive nonhodgkin's lymphoma," *N. Engl. J. Med.*, vol. 329, no. 14, pp. 987–994, 1993.
[41] D. Harrington and T. Fleming, "A class of rank test procedures for censored survival data," *Biometrika*, vol. 69, pp. 553–566, 1982.

Volker Roth was born in Klagenfurt, Austria, in 1970. He received the diploma degree in physics in 1997, and the Ph.D. degree in computer science in 2001, both from the University of Bonn, Germany.

Currently, he is a Postdoctoral Researcher with the Computer Vision and Pattern Recognition Group headed by Prof. J. M. Buhmann. His research interests include support vector machines and kernel-based learning algorithms, unsupervised learning and clustering, bioinformatics and computational biology.

Tilman Lange (S'99) was born in Bonn, Germany, in 1976. He received the diploma degree in computer science in 2002 from the University of Bonn, Germany. Currently he is a doctoral student in the Computer Vision and Pattern Recognition Group headed by Prof. J. M. Buhmann. His main research interests are in unsupervised learning and computational biology.