Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

MARCUS WEBER, WASINEE RUNGSARITYOTIN,
ALEXANDER SCHLIEP

Perron Cluster Analysis and Its Connection to Graph Partitioning for Noisy Data

ZIB-Report 04-39 (November 2004)

Perron Cluster Analysis and Its Connection to Graph Partitioning for Noisy Data

Marcus Weber[1], Wasinee Rungsarityotin[2], Alexander Schliep[2]

[1] Zuse Institute Berlin ZIB
Takustraße 7, D-14195 Berlin, Germany,
weber@zib.de
Tel: +49-30-84185-189 Fax: +49-30-84185-107

[2] Computational Molecular Biology, Max Planck Institute for Molecular Genetics
Ihnestraße 63–73, D-14195 Berlin, Germany,
rungsari | schliep@molgen.mpg.de
Tel: +49-30-8413-1166 Fax: +49-30-8413-1152

Abstract

The problem of clustering data can be formulated as a graph partitioning problem. Spectral methods for obtaining optimal solutions have reveceived a lot of attention recently. We describe Perron Cluster Cluster Analysis (PCCA) and, for the first time, establish a connection to spectral graph partitioning. We show that in our approach a clustering can be efficiently computed using a simple linear map of the eigenvector data. To deal with the prevalent problem of noisy and possibly overlapping data we introduce the minChi indicator which helps in selecting the number of clusters and confirming the existence of a partition of the data. This gives a non-probabilistic alternative to statistical mixture-models. We close with showing favorable results on the analysis of gene expression data for two different cancer types.

Keywords: Perron cluster analysis, spectral graph theory, clustering, gene expression.

MSC: 62H30, 65F15, 92-08

1 Introduction

In data analysis, it is a common first step to detect groups of data, or clusters, sharing important characteristics. The relevant body of literature with regard to methods as well as applications is vast (see [14] or [18] for an introduction). There are a number of ways to obtain a mathematical model for the data and the concept of similarity between data points, so that one can define a measure of clustering quality and design algorithms for finding a clustering maximizing this measure. The simplest, classical approach is to model data points as vectors from R^n. Euclidean distance between points measures their similarity and the average Euclidean distance between data points to the centroid of the groups they are assigned to is one natural measure for the quality of a clustering. The well-known k-means algorithm [18] will find a locally optimal solution in that setting.

One of the reasons why the development of clustering algorithms did not cease after k-means are the many intrinsic differences of data sets to be analyzed. Often the measure of similarity between data points might not fulfill all the properties of a mathematical distance function, or the measure of clustering quality has to be adapted, as for example the ball-shape assumption inherent in standard k-means does not often match the shape of clusters in real data.

An issue which is usually, and unfortunately, of little concern, is whether there is a partition of the data into a number of groups in the first place and how many possible groups the data support. Whenever we apply a clustering algorithm which computes a k-partition this is an assumption we imply to hold for the data set we analyze. The problem is more complicated when k is unknown. In the statistical literature mixture models [20] are suggested as alternatives for problem instances where groups overlap.

We address the problem of finding clusters in data sets for which we do not require the existence of a k-partition. The model we will use is a similarity graph. More specifically, we have $G = (V, E)$, where $V = \{1, \ldots, n\}$ is the set of vertices corresponding to the data points. We have an edge $\{i, j\}$ between two vertices iff we can quantify their similarity, which is denoted $w(i, j)$. The set of all edges is E and the similarities can be considered as a function $w : E \mapsto R_0^+$. The problem of finding a k-partition of the data can now be formulated as the problem of partitioning V into k subsets, $V = \cup_{i=1}^{k} V_i$. Let us consider the problem of finding a $k = 2$ partition, say $V = A \cup B$. This can be achieved by removing edges $\{i, j\}$ from E for which $i \in A$ and $j \in B$. Such a set of edges which leaves the graph disconnected is called a *cut* and the weight function allows us quantify cuts by defining their weight or *cut-value*,

$$cut(A, B) := \sum_{\{i,j\} \in E, i \in A, j \in B} w(i, j).$$

A natural objective is to find a cut of minimal value. This problem can be solved with the min-cut algorithm [22] in $O(|V||E| + |V|2log|V|)$. A problem with this objective function is that sizes of partitions do not matter. As a consequence, using min-cut will often compute very unbalanced partitions, effectively splitting V into a single vertex, or a small number of vertices, and one very large set of vertices; cf. Fig. 1. We can alleviate this problem by evaluating cuts differently.

One alternative measure is average cut, which is defined as

$$\text{Averagecut}(A, B) := \frac{cut(A, B)}{|A|} + \frac{cut(A, B)}{|B|}.$$

Average cut is sensitive to the sizes of either A and B getting too small; as long as A and B are balanced the average cut yields $2/|V|$ times the cut value, which is easily exceeded by the term for the smaller partition.

Instead of just considering partition sizes one can also consider the similarity within partitions, for which we introduce the so-called *association* value of a vertex set A denoted by $a(A) = a(A, V) := \sum_{i \in A} \sum_{j \in V} w_{ij}$. Defining the normalized cut by

$$\text{Normcut}(A, B) = \frac{cut(A, B)}{a(A, V)} + \frac{cut(A, B)}{a(B, V)},$$

we observe that the cut value is now set into relation to the similarity of each partition to the whole graph. Vertices which are more similar to many data points are harder to separate. As we will see, the normalized cut is well suited as an objective function for minimizing because it keeps the relative size and connectivity of clusters balanced.

The min-cut problem can be solved in polynomial time for $k = 2$. Finding k-way cuts in arbitrary graphs for $k > 2$ is NP-hard [6]. For the two other cut criteria, already the problem of finding a 2-way cut is NPC [Papadimitriou] [23]. However, we can find good approximate solutions [19, 23] to the 2-way normalized cut by considering a relaxation of the problem. Instead of discrete assignments to partitions consider a continuous indicator for membership. As it turns out, the eigenvectors obtained from a suitable eigenvector problem for the Laplacian of the pairwise-similarity graph G can be interpreted for exactly that purpose. This so-called spectral

method [19, 23] has been used for solving the k-partition problem directly as well as through successive computation of 2-partitions. Recall, that the weight function w is symmetric as the graph is undirected and the weights are non-negative; we will write W to denote the matrix of weights using the convention of weight zero for non-edges. With

$$d(i) = \sum_{\{i,j\} \in E} w(i,j)$$

we can define the Laplacian matrix L of G as

$$L(i,j) = \begin{cases} d_i - W(i,j) & \text{if } i = j, \\ -W(i,j) & \text{if } \{i,j\} \in E, \\ 0 & \text{else.} \end{cases} \tag{1}$$

The normal form of the Laplacian of a weighted undirected graph G is then defined to be $\mathfrak{L} = D^{-1}(D-W)$, where $D = diag(d(1),\ldots,d(n))$. For solving the average cut problem we will consider the eigenvalue problem

$$Lx = (D-W)x = \lambda x$$

and for the normalized cut problem the generalized problem

$$\begin{aligned} D\mathfrak{L}x &= \lambda Dx \\ (D-W)x &= \lambda Dx \end{aligned} \tag{2}$$

or the standard Eigenvalue problem of

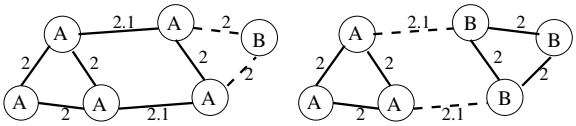
$$D^{-1/2}(D-W)D^{-1/2}x = \lambda x. \tag{3}$$

The running time of the 2-partition problem directly depends on the time to compute the second-smallest (or largest) eigenvector. For the standard eigenvalue problem $Ax = \lambda x$, it is $O(n^3)$ where n is the order of the graph. If A is sparse the Lanczos algorithm, an iterative solver, can compute an eigenvector in $O(mn) + O(mM(n))$, where m is the number of times the operation Ax is executed and $M(n)$ is the cost of matrix-vector multiplication Ax. Note that m will be typically smaller than the worst-case bound of n; the exact value depends on the sparsity. Similarly, $M(n)$ will be on the order of $O(n)$ for sufficiently sparse instances.

For solving the 2-partition problem, we are interested in the eigenvector x_2 for the second-smallest eigenvalue [19, 23]. In particular, we will inspect its sign structure and use the sign of an entry $x_2(i)$ to assign vertex i to one or the other vertex set. Similarly, for direct computation of k-partitions one can use all k eigenvectors to obtain k-dimensional indicator vectors. Previous approaches [2, 23] relied on k-means clustering of the indicator vectors to obtain a k-partition in this space.

Assume that the data can be partitioned and the assignment is given by the block diagonal matrix B. We then can write W as $B+E$, where E is some perturbation error. In this setting, under some further conditions, the recursive algorithm is guaranteed to arrive at clusters which are indeed the blocks in B.

Theorem 1.1 *[19] Let B be a perfect block diagonal matrix. Assume that the eigenvalue gap $1 - \frac{\lambda_2}{\lambda_1}$ of each block B_i, $i = 1,\ldots,k$, is at least $\beta (0 < \beta < 1)$. In addition, let the difference between the k-th and $(k+1)$-th eigenvalues of B be at least β. Let E denote the perturbation error matrix for which $W = B+E$ holds. We additionally assume E is bounded: $||E||_2 < \delta$, where $q\delta < \beta$ for some positive constant q and that the ratios of cluster sizes are bounded. Then, the recursive algorithm applied to the W matrix finds a clustering that differs from the optimal in $O(n/q^2)$ rows.*



(a) Minimum cut

(b) Normalized minimum cut

Figure 1: The first case is when minimum-cut can be wrong due to presence of low-degree vertices.

The theorem implies that if the optimal clustering has a large degree of overlapping, the spectral algorithm will fail because δ is close to β and thus the constant q will be small. With our analysis in Section 3.3 we will give an example for this. In the next section, we will propose an indicator for the amount of overlapping in W which helps in deciding whether the recursive spectral method is applicable. Subsequently we will introduce an alternative approach to finding k-partitions even in absence of a perfect block structure. We first rephrase the problem equivalently in terms of transition matrices of Markov-chains and use perturbation analysis to arrive at the main result, a geometric interpretation of the eigenvector data as a simplex. This allows to devise an assignment of data into overlapping groups and a measure for the deviation from the simplex structure, the so-called *Min-chi value*. The advantages of our method are manifold: there are fewer requirements on the similarity measure, it is effective even for high-dimensional data and foremost, with our robust diagnostic we can assess whether a unique k-partition exists. The immediate application value is two-fold. On one hand, the Min-chi value indicates whether trying to partition the data into k groups is possible. On the other hand, this can be used as a guide for deciding on the number of clusters. We demonstrate the practical effectiveness by evaluating our method on two gene expression data sets, which pose problems for partition algorithms as groups of genes sharing the same function tend to overlap and error levels as well as amount of noise tends to be very high. We conclude with a discussion with the very favorable results both for the diagnostic compared to classical cluster indices and the result of the clustering.

2 Perturbation analysis of eigenvectors

2.1 Comparing cut algorithms and the Perron Cluster Analysis

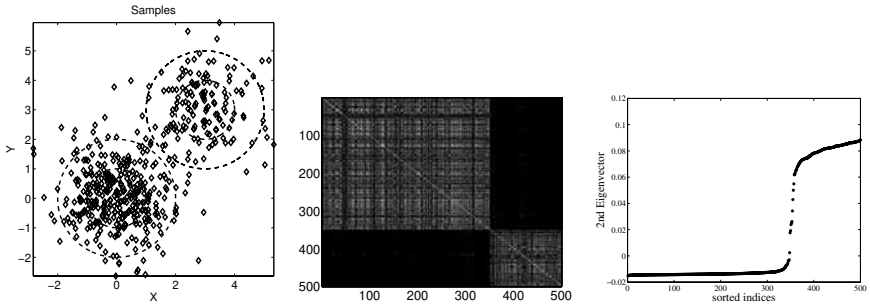
If the graph cut problem is well-defined W is nearly block structured after a suitable permutation of rows and columns. The k-way graph cut problem with an edge weight matrix W can therefore be seen as a problem of recovering the hidden block structure of W. In the ideal case, W has k pairwise unconnected vertex sets. Reweighting the rows of W via $T = D^{-1}W$ ends up in a stochastic[1] matrix T. In the ideal case, T has also block structure, where each block is a stochastic matrix, see Fig. 3 for $k = 3$.

The following equations show that the eigenvectors of T and the eigenvectors of the normalized cut problem (2) are identical:

$$\begin{aligned} & (D-W)\,y = \lambda\, D\,y \\ \Leftrightarrow\quad & D^{-1}\,(D-W)\,y = \lambda\, y \\ \Leftrightarrow\quad & y - T\,y = \lambda\, y \\ \Leftrightarrow\quad & T\,y = (1-\lambda)\,y. \end{aligned} \tag{4}$$

Because of equation (4) and $\lambda \approx 0$ the eigenvectors of T corresponding to eigenvalues $(1-\lambda) \approx 1$ are important, which are identical to the eigenvectors of the normalized cut algorithm. The ba-

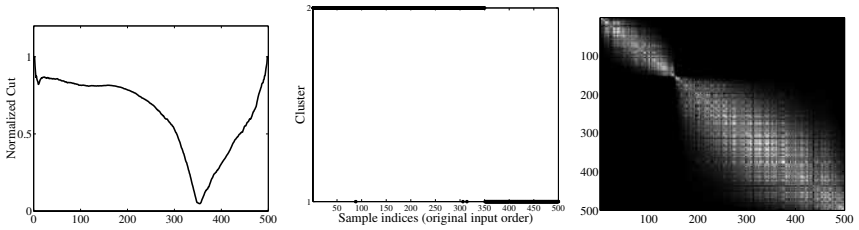
[1] T is nonnegative and the row sums are 1, i.e. $\|T\|_1 = 1$.



(a) A mixture of two standard Normal

(b) Input similarity matrix **W** (random samples from above)

(c) The sorted 2nd-Eigenvector of the Laplacian $\mathfrak{L}$



(d) Mininum of Normalized Cut

(e) Clustering by the sign-pattern

(f) Ordering of the matrix W

Figure 2: We draw 500 random samples from mixture of two normal distributions center at different means with mixing coefficients = $[0.7, 0.3]$. The entry W_{ij} is $exp(-||x_i - x_j||^2)$ for all i, j. This picture shows that in the case of no overlapping of two clusters, the minimum of the normalized cut occurs at the zero-crossing.

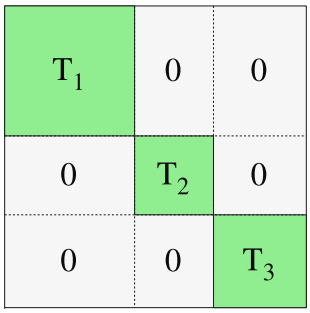


Figure 3: Structure of the stochastic matrix T in the case of $k = 3$ pairwise unconnected vertex sets after a suitable permutation of row and column indices.

sic idea of the Perron (Cluster) Cluster Analysis (PCCA) is, that these eigenvectors are almost invariant under T, and that one can apply Markov chain theory to the stochastic matrix T, see Deuflhard et al. [7, 8]. The name PCCA derives from the so-called Perron eigenvalue $\lambda_1 = 1$ of a stochastic matrix [4, 21] and from the fact, that one uses eigenvectors corresponding to a cluster of eigenvalues near λ_1 for clustering the data.

In the case, where W is symmetric, in [7] it is shown that an ε-perturbation $\widetilde{T}$ of the stochastic matrix T, where

$$\widetilde{T} = T + \varepsilon T^{(1)} + O(\varepsilon^2), \tag{5}$$

leads to an ε-perturbation of the components of the corresponding eigenvectors $\widetilde{y} = y + O(\varepsilon)$. Furthermore it is shown, that in the ideal case the sign structure of the k observed eigenvectors determines the index sets of the k uncoupled blocks uniquely.

This leads to a former version of PCCA, where the sign structure of the perturbed eigenvectors is examined in order to find the hidden blocks of $\widetilde{T}$. For $k = 2$, we now show that the former PCCA method and the 2-way partition method previously mentioned [19, 23] are the same. However both methods are not robust because the sign information may be sensitive to slight perturbations (e.g. "dirty zero" problem). In practise, the generalized eigenvalue problem for the normalized cut algorithm is often replaced by the solution of a symmetric problem, see equation (3). The following equations show, that $y = D^{-1/2}x$ is just a positive component-wise scaling of the eigenvector x and therefore the eigenvectors $x_1, \ldots, x_k$ and $y_1, \ldots, y_k$ corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_k$ share the same sign structure.

$$\begin{aligned} & D^{-1/2}(D-W)D^{-1/2}x = \lambda x \\ \Leftrightarrow \quad & D^{-1/2}(D-W)y = \lambda D^{1/2}y \\ \Leftrightarrow \quad & (D-W)y = \lambda D y \\ \Leftrightarrow \quad & Ty = (1-\lambda)y. \end{aligned} \tag{6}$$

Equation (6) also shows that the spectrum of T is real valued for a symmetric weight matrix W. In Section 2.2 it is shown, that there is a robust version of PCCA using not only the sign structre of the eigenvectors, but also the values of their components.

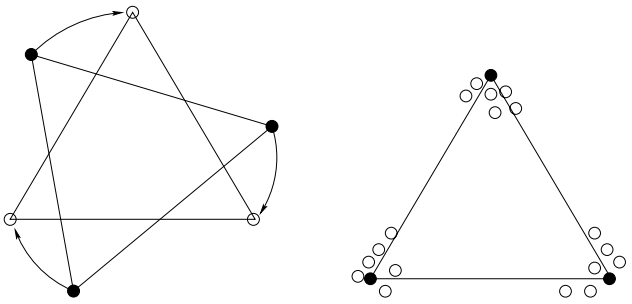
2.2 Second order perturbation result - Simplex structure

From a different point of view Huisinga et al. [16] examine the correspondence between the sign structure of eigenfunctions of a Markov operator (the continuos version of a stochastic matrix) and the characterization of so-called transition sets. From their context it seems natural, that in the general perturbed case the k first eigenvectors $\widetilde{Y} = [\widetilde{y}_1, \ldots, \widetilde{y}_k]$ of $\widetilde{T}$ comprise all 2^{k-1} possible[2] sign structures regarding the rows of $\widetilde{Y}$. For the former PCCA algorithm, this means

[2]Note, that $\widetilde{y}_1$ is a constant vector.

that for $k > 2$ there are more sign structures than clusters. Deuflhard have shown, that in this case a so-called "dirty zero" problem occurs from perturbation of the sign "0", see [7, 8]. Deuflhard and Weber [8] overcome this problem by examing a second order perturbation result and by using a mixture model for the clustering, which leads them to the Robust Perron Cluster Analysis (PCCA+). Here we recall some of their ideas.

In the ideal uncoupled case the components of each of the eigenvectors $Y = [y_1, \ldots, y_k]$ of T are pairwise identical for indices corresponding to the same block. Regarding the rows of Y as points in $\mathbb{R}^k$, therefore, ends up in exactly k different points. Since the first eigenvector y_1 corresponding to $\lambda_1 = 1$ for a stochastic matrix is always constant, this can also be seen as k different points in $\mathbb{R}^{k-1}$ omitting the first column of Y. The convex hull of k different points in $\mathbb{R}^{k-1}$ is called a *simplex*. In $\mathbb{R}^2$ for $k = 3$ it is a triangle, see Fig. 4. In $\mathbb{R}^3$ for $k = 4$ it is a tetrahedron.



(a) First order perturbation

(b) Higher order perturbation

Figure 4: Perturbation result of Deuflhard and Weber, Lemma 1.1. in [8], for stochastic matrices with symmetric W: The simplex structure of the eigenvectors is saved in first order perturbation theory.

With the ε-perturbation analysis of equation (5) the first order perturbation term for the eigenvectors

$$\widetilde{Y} = Y + \varepsilon Y^{(1)} + O(\varepsilon^2)$$

is a linear mapping $Y^{(1)} = YB$, with a regular matrix $B \in \mathbb{R}^{k \times k}$, of the ideal case eigenvectors, see Fig. 4(a). The simplex structure is not perturbed. The perturbation of the simplex structure Y vs. $\widetilde{Y}$ is therefore at most of order $O(\varepsilon^2)$, see Fig. 4(b). This simplex structure plays an important role in the PCCA+ algorithm.

For the examination of gene expression data often the so-called fuzzy clustering methods yield good results [11, 12, 13]. In fuzzy clustering for every cluster $i = 1, \ldots, k$ there is a membership function $\widetilde{\chi}_i : V \to [0, 1]$, which assigns a grade of membership between 0 and 1 to each vertex (gene pattern) in V. Therefore, a vertex may correspond to different clusters with a different grade of membership. For each vertex $v \in V$ the sum of the grades of membership with regard to the different clusters is 1, i.e.

$$\sum_{i=1}^{k} \widetilde{\chi}_i(v) = 1. \tag{7}$$

Each vertex is represented by a fuzzy vector $(\widetilde{\chi}_1(v), \ldots, \widetilde{\chi}_k(v))' \in \mathbb{R}^k$. Since these vectors are positive and the partition of unity holds, they lie in the standard σ_{k-1} simplex spanned by the k unit vectors of $\mathbb{R}^k$. Note, that (7) defines a $(k-1)$-dimensional subspace in $\mathbb{R}^k$. At this stage: The result of fuzzy clustering is a simplex and the input data set (the rows of the eigenvector matrix $\widetilde{Y}$ of $\widetilde{T}$) has also an $O(\varepsilon^2)$-perturbed simplex structure. Therefore, clustering can be seen

as a simple linear mapping from the rows of $\widetilde{Y}$ to the rows of a membership matrix $\widetilde{\chi}$, where $\widetilde{\chi}_{ji} := \widetilde{\chi}_i(v_j)$ and $j = 1,\ldots,N$ is the index of the vertex $v_j \in V$. For the unperturbed ideal case, where T has block structure, this linear mapping can be done such that the membership matrix χ comprises the k characteristic functions for the clusters. The fact, that the mapping between the perturbed eigenvectors $\widetilde{Y}$ and the membership matrix $\widetilde{\chi}$ is linear and deviation from simplex structure is $O(\varepsilon^2)$ leads to the following estimation

$$\chi - \widetilde{\chi} = O(\varepsilon^2),$$

which is the key equation showing the robustness of the method in [8], it shows a bounded convergence to the strict clustering as $\varepsilon \to 0$.

3 Recovering the simplex structure of eigenvectors

3.1 Construction of a linear mapping

The linear mapping is expressed by a regular $k \times k$-matrix $\mathscr{A}$:

$$\widetilde{\chi} = \widetilde{Y}\mathscr{A}. \tag{8}$$

This matrix maps the vertices of the simplex contained in the rows of $\widetilde{Y}$ onto the vertices of the simplex σ_{k-1}, i.e. the k unit vectors. Therefore, if one finds the indices $\pi_1,\ldots,\pi_k \in \{1,\ldots,N\}$ of the vertices in $\widetilde{Y}$ one can construct the linear mapping via

$$\mathscr{A}^{-1} = \begin{pmatrix} \widetilde{Y}_{\pi_1,1} & \cdots & \widetilde{Y}_{\pi_1,k} \\ \vdots & & \vdots \\ \widetilde{Y}_{\pi_k,1} & \cdots & \widetilde{Y}_{\pi_k,k} \end{pmatrix}. \tag{9}$$

Equation (9) leads to a feasible solution $\widetilde{\chi} = \widetilde{Y}\mathscr{A}$ if and only if the convex hull of the rows of $\widetilde{Y}$ (seen as vectors in $\mathbb{R}^k$) is a simplex [27]. From perturbation analysis we know that this is the case with a deviation of order $O(\varepsilon^2)$. In the general case, the result obtained by (9) is infeasible.

However, the partition of unity constraint is always satisfied, because the rows of $\widetilde{Y}$ as vectors in $\mathbb{R}^k$ lie in a $(k-1)$-dimensional subspace, which is mapped to the subspace defined by (7). Therefore, deviation from simplex structure leads to infeasibility of the linear mapping only with regard to the bounds $0 \leq \widetilde{\chi}_{ji} \leq 1$, especially to some negative grades of membership. For more details see Theorem 2.1 in [8]. There are two possibilities to cope with this situation:

- In [8] the linear mapping is modified, such that the solution is feasible and optimal with regard to some cost function.
- The linear mapping is not modified. The lowest value in the result matrix $\widetilde{\chi}$, the so called *minChi-value*, is used as indicator for the deviation from simplex structure and should be $O(\varepsilon^2)$. We prefer this possibility in the present text.

Huisinga et al. [15, 17] and Deuflhard et al. [8] have defined a measure of metastability and they have shown that the metastability of k subsets of a Markovian system can be bounded from above by the sum of the first k eigenvalues of the corresponding stochastic matrix T. Therefore, in the context of Markov chain theory the eigenvalues of T can be used as an indicator for the number of clusters, i.e. only eigenvalues near the Perron root $\lambda_1 = 1$ should be taken into account. In the present paper, however, the stochastic matrix T is not computed from a Markov chain, but from a geometrical clustering. In computer experiments, the second eigenvalue of T was always bounded away from $\lambda_1 = 1$, but the simplex structure can still be found. Therefore, the metastability indicator or the eigenvalue criterion does not hold for this type of clustering any more. Also the optimization of metastability like in [8] does not make sense for a geometrical probelm. This is the reason to apply the minChi-value described before as indicator for the number of clusters.

The relation between the simplex structure of the eigenvectors of a stochastic matrix $T = D^{-1}W$ and the fuzzy clustering idea has first been examined by Weber and Galliat [27]. Their "inner simplex algorithm" can be used to find the indices $\pi_1, \ldots, \pi_k$ of the rows of $\widetilde{Y}$, which are used for the construction of $\mathscr{A}$. The computational effort of this algorithm is of order $O(N^2)$, where N is the number of vertices of the graph G. An improvement of this algorithm, which has a computational effort of order $O(N)$, is given in [8], where the inner simplex algorithm leads to a starting guess for the optimization routine. If we denote the rows of $\widetilde{Y}$ by $\widetilde{Y}(1), \ldots, \widetilde{Y}(N) \in \mathbb{R}^k$, then the algorithm, which determines the index set $\{\pi_1, \ldots, \pi_k\}$ for (9), is as follows:

Inner Simplex Algorithm

1. Define π_1 as that index, for which $\|\widetilde{Y}(\pi_1)\|_2$ is maximal. Define $\mathscr{X}_1 = \text{span}\{\widetilde{Y}(\pi_1)\}$.
2. For $i = 2, \ldots, k$: Define π_i as that index, for which the distance to the hyperplane $\mathscr{X}_{i-1}$, i.e. $\|\widetilde{Y}(\pi_i) - \mathscr{X}_{i-1}\|_2$, is maximal. Define $\mathscr{X}_i = \text{span}\{\widetilde{Y}(\pi_1), \ldots, \widetilde{Y}(\pi_i)\}$. Go on with 2.

3.2 A guiding example

First we construct a stochastic matrix $\widetilde{T}$. For this reason we take a symmetric 6×6-matrix W having perfect block structure and perturb this matrix with some perturbation terms. The result matrix is:

$$\widetilde{T} = \begin{pmatrix} 0.3432 & 0.1663 & 0.1367 & 0.1377 & 0.1085 & 0.1076 \\ 0.1672 & 0.3427 & 0.1370 & 0.1377 & 0.1080 & 0.1074 \\ 0.1078 & 0.1075 & 0.3222 & 0.2476 & 0.1073 & 0.1075 \\ 0.1083 & 0.1077 & 0.2470 & 0.3216 & 0.1081 & 0.1073 \\ 0.1086 & 0.1076 & 0.1363 & 0.1377 & 0.3435 & 0.1663 \\ 0.1080 & 0.1073 & 0.1369 & 0.1370 & 0.1667 & 0.3443 \end{pmatrix}.$$

Clustering. The eigenvalues of $\widetilde{T}$ are $\{1.0000, 0.2953, 0.2940, 0.1774, 0.1762, 0.0746\}$. The eigenvalue indicator for the number of clusters gives us $k = 1$, but we know by construction $k = 3$. In the following we will examine the cases k=3, k=4, and k=2. The first 4 eigenvectors of $\widetilde{T}$ can be calculated as:

$$\widetilde{Y} = \begin{pmatrix} 1.0000 & 0.1079 & -1.4997 & 0.2324 \\ 1.0000 & 0.0986 & -1.5087 & -0.2480 \\ 1.0000 & -1.1116 & 0.5921 & -0.0097 \\ 1.0000 & -1.1014 & 0.5856 & 0.0087 \\ 1.0000 & 1.2989 & 0.7447 & 1.7987 \\ 1.0000 & 1.3115 & 0.7641 & -1.7873 \end{pmatrix}.$$

Clustering, k=3. With the above algorithm we get $\pi_1 = 6, \pi_2 = 2$, and $\pi_3 = 3$. Therefore,

$$\mathscr{A}^{-1} = \begin{pmatrix} 1.0000 & 1.3115 & 0.7641 \\ 1.0000 & 0.0986 & -1.5087 \\ 1.0000 & -1.1116 & 0.5921 \end{pmatrix}, \quad \mathscr{A} = \begin{pmatrix} 0.3055 & 0.3068 & 0.3877 \\ 0.3965 & 0.0325 & -0.4289 \\ 0.2284 & -0.4573 & 0.2289 \end{pmatrix}.$$

Applying the transformation $\widetilde{\chi} = \widetilde{Y}\mathscr{A}$, where $\widetilde{Y}$ comprises the first 3 eigenvectors, yields

$$\widetilde{\chi} = \begin{pmatrix} 0.0057 & 0.9962 & -0.0019 \\ 0.0000 & 1.0000 & 0.0000 \\ 0.0000 & 0.0000 & 1.0000 \\ 0.0026 & 0.0033 & 0.9941 \\ 0.9906 & 0.0085 & 0.0010 \\ 1.0000 & 0.0000 & 0.0000 \end{pmatrix},$$

which is almost feasible, because minChi $= -0.0019 \approx 0$. The 3 columns of $\widetilde{\chi}$ are the corresponding membership functions, which separate the 3 blocks of W well.

Clustering, k=4. With the above algorithm we get $\pi_1 = 6, \pi_2 = 5, \pi_3 = 2$, and $\pi_4 = 3$. Computing $\mathscr{A}$ by equation (9) and using $\widetilde{\chi} = \widetilde{Y}\mathscr{A}$ yields:

$$\widetilde{\chi} = \begin{pmatrix} -0.1301 & 0.1371 & 0.9950 & -0.0021 \\ 0.0000 & 0.0000 & 1.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 1.0000 \\ 0.0040 & 0.0066 & 0.0032 & 0.9941 \\ 0.0000 & 1.0000 & 0.0000 & 0.0000 \\ 1.0000 & 0.0000 & 0.0000 & 0.0000 \end{pmatrix},$$

which is infeasible, because minChi $= -0.1301 \ll 0$ is about 100 times the value of minChi for $k = 3$.

Clustering, k=2. For the case $k = 2$ we always get minChi $= 0$, see [8, 27]. The minChi-indicator prefers this number of clusters. It is easy to decide, whether the right number of clusters is $k > 2$, if we take the maximum number $k \ll N$ for which minChi ≈ 0 (like in the present example). But if minChi $\ll 0$ for every $N \gg k > 2$, then the decision, whether $k = 1$ or $k = 2$ is the right number of clusters is not easy. In this case, one can e.g. look at the result matrix $\widetilde{\chi}$ for $k = 2$ and determine how well the 2 clusters are separated, i.e. if $\widetilde{\chi}_{ij} \approx 0$ or $\widetilde{\chi}_{ij} \approx 1$ for every $i = 1, \ldots, N$ and $j = 1, 2$ (in other words: if the information entropy is low, see [25]).

3.3 Less input vectors lead to a wrong clustering

In spectral k-partitioning methods often a smaller number s of eigenvectors, e.g. $s = \lceil \log_2 k \rceil$ in [9, 10], is used as input data for cluster algorithms. A counter-example following the ideas of Weber [26] shows, that the projection into lower subspaces might cause failures in classification. If we have "full" simplices, see Fig. 5 top, a projection of the simplex structure into a lower dimensional subspace conceals the difference between transition states[3] ($A \leftrightarrow B$) and vertices corresponding to a different cluster (C), see Fig. 5 bottom. Therefore, especially successive algorithms working only on the eigenvector corresponding to the 2nd largest (or 2nd lowest for the Laplacian) eigenvalue might fail [19, 23]. Only with strong assumptions like in Theorem 1.1, correctness can be shown. In the example figure, the marked points are assigned to cluster C, if one uses only 2 eigenvectors. But in PCCA the 3rd eigenvector shows, that one of the marked points do not correspond to C but is a transition state between A and B and should be assigned to one of those clusters.

4 minChi in practice

Given an $n \times m$ data matrix, we compute pairwise-distances for all pairs and construct the $n \times n$ distance matrix A with a symmetric distance function $w : R^m \times R^m \mapsto R_0^+$. We then convert the distance to a similarity matrix with $W = \exp(-\beta A)$ where β is a scaling parameter and the stochastic matrix is defined by $T = D^{-1}W$. We can use the error measure min-Chi to determine a locally optimal solution for the number of clusters. Given the matrix T, we can use our method to determine a number of clusters denoted by k as follows:

The Mode Selection Algorithm

1. Choose $k_{min} \ldots k_{max}$ such that the optimal k could be in the interval,

[3] Transition states are vertices of the weighted graph, which have a similar grade of membership with regard to different clusters.

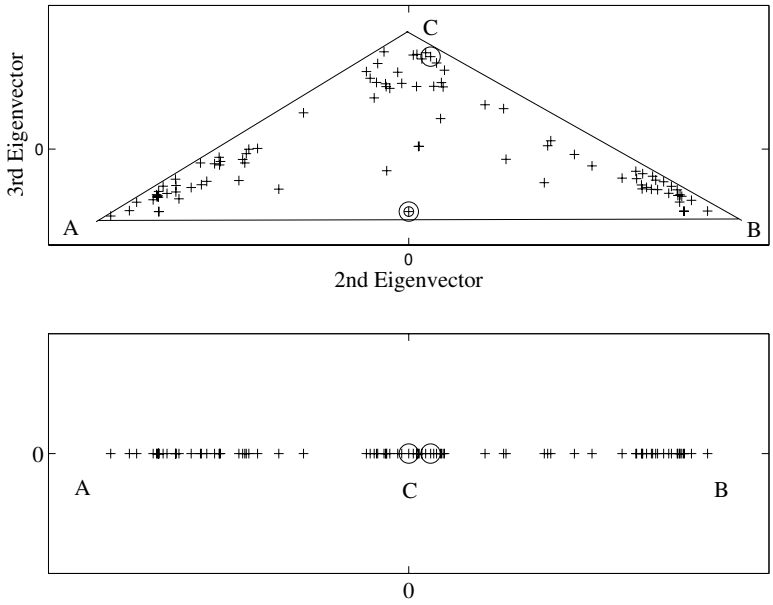


Figure 5: For a 3-clustering the PCCA method also needs 3 eigenvectors (see top figure, 1st eigenvector is constant). If one uses only the 2nd eigenvector of T for a 3-clustering, important information about membership is lost (see botom figure and the marked points)

2. Iterate from $k_{min} \ldots k_{max}$ and for each k-th trial, calculate χ (eq. 8) for cluster assignment via the *Inner Simplex* algorithm and min-Chi as an indicator for the number of clusters,
3. Choose the maximum k for which min-Chi $<$ Thresholds as the number of clusters.

Selections of the threshold depends on the value β which controls the perturbation from the perfect block structure of T (see Figure 9). As a rule, when β is large, the threshold can be small because T is almost block-diagonal.

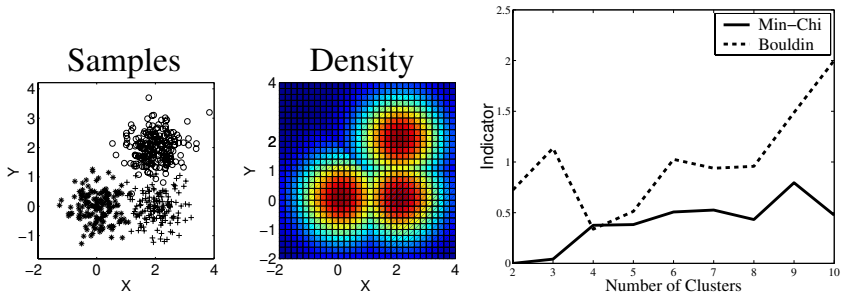
We argue for the suitability of the Min-Chi indicator in the case of the overlapping clusters because traditional internal indices are not applicable when the average distance within-cluster is much larger than the distances between the centers. The minimum separation of the means is the assumption used to define internal indices such as the Bouldin index which is not suitable for a model of overlapping clusters. We compare the Min-Chi indicator with the Bouldin index applied to the result from the Inner Simplex algorithm 3.1. We simulate two data sets distributed as a mixture of bivariate Normal. In Fig. 6, we show the result for the mixture of four Normal distributions with one of high variances and in Fig. 6 the results for the mixture of three standard Normal distributions.

Given a partition into K clusters by a clustering algorithm, one first defines the measure of within-to-between cluster spread for the kth cluster with the notation $R_k = \max_{j \neq k} \frac{e_j + e_k}{m_{jk}}$, where e_i is the average distance within the ith cluster and m_{ij} is the euclidean distance between the means. The Bouldin index [18] for k is

$$DB(K) = \frac{1}{K} \sum_{k>1} R_k.$$

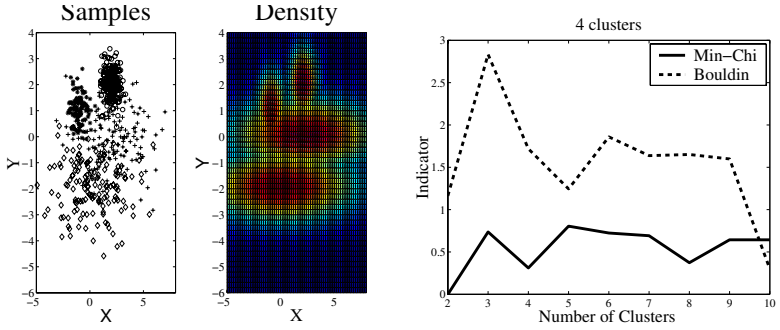
According to the Bouldin indicator, the number of clusters is k^* such that

$$k^* = \underset{k_{min} \ldots k_{max}}{\operatorname{argmin}} \; DB(k).$$



(a) Samples and their density (b) Min-Chi and Bouldin Indicator

Figure 6: **Simulated data (s1):** mixture of three spherical gaussians, where the distance between the means is greater than the variance.



(a) Samples and their density (b) Min-Chi and Bouldin Indicator

Figure 7: **Simulated data (s2):** mixture of four gaussians, where the minimum separation between the means does not hold.

We calculate both the Min-Chi indicator and the Bouldin index for $k = \{2, \ldots, 10\}$ and illustrate the result in Fig. 7 and 6.

4.1 Example: clustering of gene expression data

Microarrays can be used to quantify the expression of messenger RNA (mRNA) of a large set of genes simultaneously. The experiments are repeated under different conditions, for different subpopulations or for different tissues. For each experiment a high-dimensional vector, this might be on the order of 30,000 dimensions for Humans, of per-gene expression levels is obtained. Clearly, the number of experiments and hence the number of data points will typically be substantially smaller. Possible goals of performing the experiments are inference of genetic variations implicated in diseases or identification of disease sub-types.

It is well known, that for the same amount of data clustering is more difficult in high-dimensional spaces. Even for a simple model such as a multi-variate Gaussian the number of parameters describing is quadratic in the dimension of the input space for unconstrained covariance matrices. Countering the need for more data to obtain reliable parameter estimates with simplifying assumptions about the covariance structure — e.g., assuming spherical Gaussians — might lead to poor representations of the data.

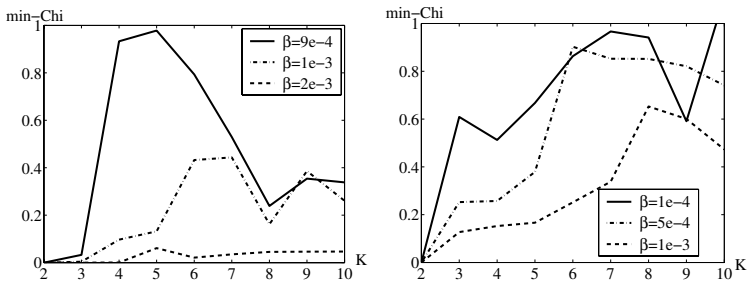
Reducing the dimension of the data prior to clustering is one classical way of tackling this problem. One approach of dimensionality reduction is to simply pick the subset of genes that is particularly meaningful. Of course, making the decision which subset to pick is difficult, because they are combinatorially many. One such system following this approach, CLIFF [28], is using information theory to rank genes, picking the ones with highest information gain. Our work is similar to the work by [2] with *explicit use of simplex structure to determine the number of clusters and compute the assignment of samples to clusters based on their positions in the simplex*. The method is an unsupervised clustering because we do not pick the subset of genes in advance. In practice, the approach should be combined with many cross-validation trials in which the number of genes varies.

4.2 Result and Discussion

Our goal is to evaluate the method on its ability to estimate the number of clusters with respect to (1) the simplex structure as a model for clustering in Section 2.2 and (2) the expert's classification on diseased subtypes and survival study.

Selection of k. Because selection of k depends on parameters β and the threshold on min-Chi, we should perform a search over the parameter set to obtain multiple candidates for optimal k. For example, the effect of scale β on the block-structure of T is shown in Figure 9. Given our model, $k = 2$ is always possible unless the data set contains no clusters ($k = 1$) in which case we need to inspect the entropy of χ to decide between partitioning into one or two clusters. By plotting k vs. min-Chi as β increases (Figure 8), we observe that the threshold on min-Chi decreases as β increases in all experiments. When β is small, there is more perturbation off diagonal and β is large, we observe that certain blocks become more distinct, but that singletons also occur.

Evaluation of cluster assignment χ. As we compute min-Chi from χ, for each possible candidate for k we also obtain a grade of membership χ and use it to assign samples to a particular cluster simply by choosing the maximum $C_i = \underset{k}{argmax}\chi_{ik}$. In theory if the projection of all points lies perfectly within the convex hull of a standard k-simplex, one can use information theory approach by applying a cut-off on the entropy of χ to decide for which cluster. However, due to numerical errors, negative weights can occur and constrained optimization must be performed but it is beyond the scope of this paper. In both experiments, we ignore the case when infeasible solutions of χ occurs and regard negative weights as zero. After we decide on cluster



(a) Min-chi indicator for Bittner (b) Min-chi indicator for Alizadeh

Figure 8: **Parameters and min-Chi.** The behavior of min-Chi depends on β and selection of k depends on the threshold of min-Chi values. In both figures, legends show min-Chi plots and their β.

membership, the difference between clusters and true classes is measured by all-pairs sensitivity and specificity which are shown in Tables 1(a) and 2(a). Specificity is a fraction of true positives over all positive classification: $\frac{\#TP}{(\#TP+\#FP)}$. Sensitivity is how many true positives identified over all pairs of samples: $\frac{\#TP}{(\#TP+\#FN)}$. A *true positive* (TP) counts when a pair of samples (s_i, s_j) is assigned to the same cluster and belongs to the same class. A *false positive* (FP) counts when (s_i, s_j) is assigned to the same cluster, but does not belong to the same class. A *false negative* (FN) counts when (s_i, s_j) is not assigned to the same cluster, but belongs to the same class.

To analyze correspondence with cancer subtypes, we evaluate our method on two gene expression data sets: Alizadeh [1] and Bittner [5], posing problems because groups of genes sharing the same function tend to overlap and noise tends to be high. For preprocessing, we follow the normalization procedure in [24], select only 2000 genes with the highest variance and use the standard Euclidean distance measure, but first correct for the mean and variance of the data set.

The second study is to find the distinction of the survival rate of patients among clusters. We use the same comparison method as in [3] on three data sets: **AML, Breast cancer, DLBCL**. In this experiment, we select only 2000 genes with the highest variances, compute two clusters using the Inner Simplex algorithm, assign patients based on the cut-off on χ ($\chi > 0.6$) and compare the difference in the distribution of survival time between the two clusters, using the Survival package implemented in **R**. We show the result for three data sets in Figure 12, 13(a),and 13(b). The p-value for each data sets and comparison with other clustering methods is summarized in Table 3. More details on the three data sets and other methods are explained in [3]. Our result suggests that even without a model for gene selection, the clusters discovered are significant with respect to the survival model and the p-value should be lower with a model for gene selection.

5 Conclusion

In this paper we have shown the relation between Perron Cluster Cluster Analysis and spectral clustering methods. Some changes of PCCA+ with regard to geometrical clustering have been proposed, e.g. the minChi indicator for the number k of clusters. We have shown that this indicator is valueable also for noisy data. It evaluates the deviation of some eigenvector data from simplex structure and, therefore, it indicates the possibility of a "fuzzy" clustering, i.e. a clustering with a certain number of almost characteristic functions. A simple linear mapping of the eigenvector data has to be performed in order to compute these almost characteristic functions. Therefore, the cluster algorithm is easy to implement and fast in practice. We have also shown, that PCCA+ does not need strong assumptions like other spectral graph partitioning methods,

Input Parameters			Result		
β	λ_2	Threshold	k	Sens.	Spec.
1e-4	0.1075	—	1	—	—
5e-4	0.533	0.0	2	1.0	0.48
		0.3	4	0.58	0.58
		0.4	5	0.58	0.79
1e-3	0.8635	0.0	2	1.0	0.48
		0.15	4	0.89	0.54
		0.2	5	0.56	0.61

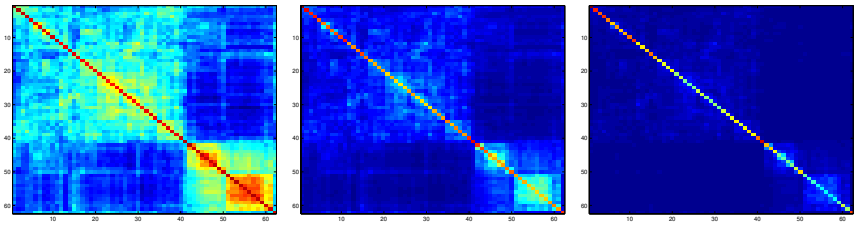
(a) Number of clusters, Sens. & Spec.

Table 1: **Alizadeh:** Summary of parameters and their effect on the number of clusters k. The feature set consists of 2000 genes used to compute the stochastic matrix in Fig. 9. The standard used to compute all-pair sens. and spec. is the expert's classification given 4 known subtypes. For β=1e-4, there is no possible mapping onto any k-simplex for $k > 1$ which indicates that with the Euclidean distance at this scale the block structure does not exist.

Input Parameters			Result		
β	λ_2	Threshold	k	Sens.	Spec.
9e-4	0.4329	0.0	2	0.92	0.53
		0.1	3	0.91	0.53
		0.3	8	0.40	0.85
1e-3	0.5050	0.0	2	0.92	0.53
		0.05	3	0.91	0.53
		0.2	8	0.38	0.77
2e-3	0.9564	0.0	2	0.92	0.53
		0.01	4	0.81	0.67
		0.03	6	0.77	0.71

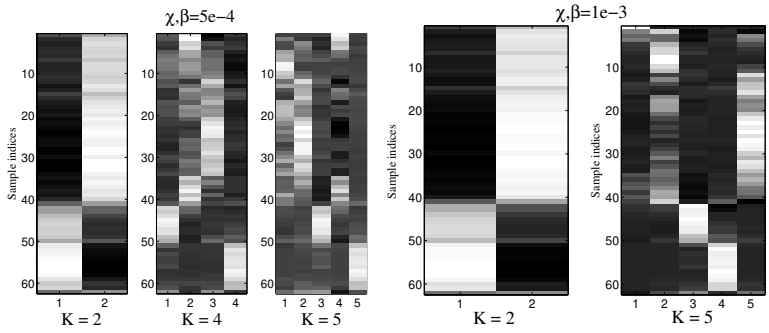
(a) Number of clusters, Sens. & Spec.

Table 2: **Bittner:** Summary of parameters and their effect on the number of clusters k. The feature set consists of 2000 genes used to compute the stochastic matrix (not the same as in the Alizadeh's set). The standard used to compute all-pair sens. and spec. is the expert's classification given 2 known subtypes.



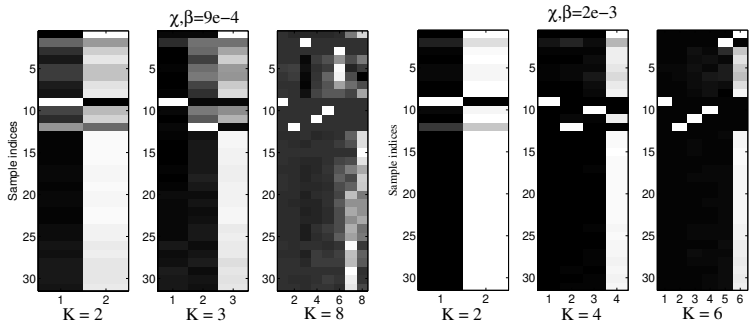
(a) scale = 1e-4 (b) scale = 5e-4 (c) scale = 1e-3

Figure 9: **Alizadeh:** Effect of different scales on the block structure. The scale parameter, β controls the amount of perturbation from the block structure of T.



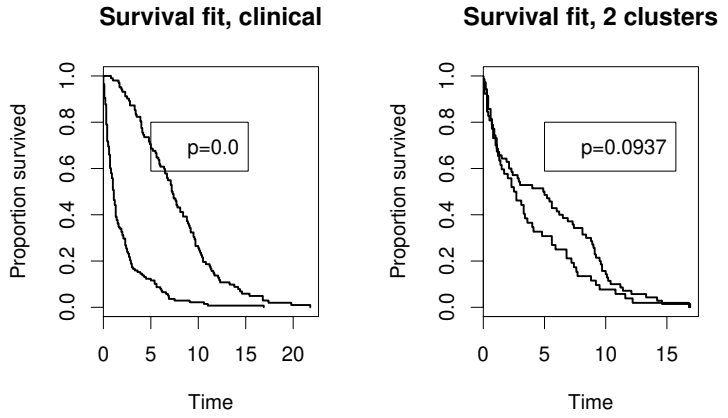
(a) **Ali,**$\beta = 5e-4$ (b) **Ali,**$\beta = 1e-3$

Figure 10: χ of Alizadeh with scale parameter $\beta = 5e-4, 1e-3$. In the figures, the gray scale maps to the interval $[0.0, 1.0]$ and each row sums to 1.0.



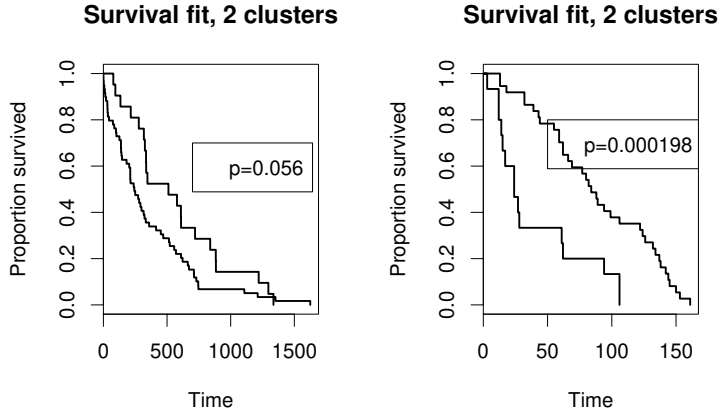
(a) **Bittner,**$\beta = 9e-4$ (b) **Bittner,**$\beta = 2e-3$

Figure 11: χ of Bittner with scale parameter $\beta = 9e-4, 2e-3$. In the figures, the gray scale maps to the interval $[0.0, 1.0]$ and each row sums to 1.0.



(a) **DLBCL**: clinical study (b) **DLBCL**: Inner Simplex clustering

Figure 12: Comparison of the Survival Curves resulting from applying the Inner Simplex algorithm on the **DLBCL** data and the actual clinical study. Our p-value is 0.0937.



(a) **AML**: two clusters (b) **Breast cancer**: two clusters

Figure 13: Comparison of the Survival Curves resulting from applying the Inner Simplex algorithm on **AML** and **Breast cancer** data sets.

Method	**DLBCL**	**AML**	**Breast cancer**
	p-value	p-value	p-value
(1) Simplex	0.0937	0.056	0.000198
(2) Von Heydebreck et. al	0.645	n/a	n/a
(3) Median Cut	0.0297	0.0487	0.00423
(4) Clustering Cox	0.00755	0.0309	0.000127
(5) Supervised PC	0.00124	0.00136	2.06e-4
(6) Clustering-PLS	0.00087	0.050	0.00123

(a) **Survival analysis**: comparison of the different methods on three data sets

Table 3: Comparison of the different methods applied to the DLBCL data of Rosenwald et al., the acute myeloid leukemia (AML) and the breast cancer data of van't Veer et al. The methods are (1) Inner Simplex algorithm; (2) using the method of [24]; (3) assigning samples to a low-risk or high-risk group based on their median survival time; (4) using 2-means clustering based on the genes with the largest Cox scores; (5) using the supervised PCA method; (6) using 2-means clustering based on the genes with the largest PLS-corrected Cox scores. Lower p-values from methods (3)-(6) is due to applying gene selections before clustering.

because it uses the full eigenvector information and not only signs or less than k eigenvectors. In the examples we transformed the almost characteristic functions back into a crisp clustering. The philosophy of robust Perron Cluster Analysis, however, tells us, that real transition states might occur, which we can not assign to only one of the clusters. The meaning of this philosophy in the field of gene expression may be a subject for further investigations.

Acknowledgements. The authors would like to thank Anja von Heydebreck for the normalization of the first two data sets. We thank Peter Deuflhard and Martin Vingron for providing additional support. Wasinee Rungsarityotin is supported by the Max Planck Society fellowship for international graduate students.

References

[1] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. r. Hudson J, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–11, Feb 2000.

[2] Andrew Y. Ng, Michael I. Jordan and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems 14*, 2002.

[3] E. Bair and R. Tibshirani. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol*, 2(4):E108, Apr 2004.

[4] A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. Academic Press, New York, 1979. Reprinted by SIAM, Philadelphia, 1994.

[5] 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*, 406(6795):536–40, Aug 2000.

[6] E. Dahlhaus, D. S. Johnson, C. H. Papadimitriou, P. D. Seymour, and M. Yannakakis. The complexity of multiterminal cuts. *SIAM J. Comput.*, 23(4):864–894, 1994.

[7] P. Deuflhard, W. Huisinga, A. Fischer, and Ch. Schütte. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Lin. Alg. Appl.*, 315:39–59, 2000.

[8] P. Deuflhard and M. Weber. Robust Perron Cluster Analysis in Conformation Dynamics. Technical Report ZIB 03-19, Zuse Institute Berlin, 2003. To appear in: J. Lin. Alg. App., 2004.

[9] Inderjit S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. *Dept. of Computer Sciences, University of Texas, Austin, TX78712*, 2001.

[10] G. Froyland and M. Dellnitz. Detecting and locating almost-invariant set and cycles. *Dept. of Mathematics and Statistics, University Paderborn*, 2001.

[11] L. M. Fu and E. Medico. FCM, a fuzzy map clustering algorithm for microarray data analysis. Bioinformatics ITalian Society Meeting (BITS04), Padova, March 2004.

[12] A. P. Gasch and M. B. Eisen. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biology*, 3(11):research0059.1–0059.22, 2002.

[13] R. Gunthke, D. Hahn, B. Fahnert, T. Kroll, and S. Wölfl. Gene expression data mining by fuzzy-c-means clustering and fuzzy rule generation. 11^{th} International Biotechnology Symposium, Berlin, September 2000.

[14] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, Berlin, 2001.

[15] W. Huisinga. *Metastability of Markovian Systems: A transfer operator based approach in application to molecular dynamics*. PhD thesis, Free University Berlin, 2001.

[16] W. Huisinga, S. Meyn, and Ch. Schütte. Phase transitions and metastability in markovian and molecular systems. *Ann. Appl. Probab.*, pages 419–458, 2004.

[17] W. Huisinga and B. Schmidt. Metastability and dominant eigenvalues of transfer operators. *Preprint, Free University Berlin*, June 2004.

[18] Anil Jain and Richard Dubes. *Algorithms for Clustering Data*. Prentice-Hall, 1988.

[19] Ravi Kannan, Santosh Vempala, and Adrian Vetta. On Clusterings: Good, Bad and Spectral. *Proceedings of IEEE Foundations of Computer Science*, 1999.

[20] G.J. McLachlan and K.E. Basford. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, Inc., New York, Basel, 1988.

[21] R.B. Bapat and T.E.S. Raghavan. *Nonegative Matrices and Applications*. Cambridge University Press, 1997.

[22] R. Shamir and R. Sharan. CLICK: a clustering algorithm with applications to gene expression analysis. *ISMB*, pages 307–316, 2000.

[23] Jianbo Shi and Jitendra Malik. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[24] A. von Heydebreck, W. Huber, A. Poustka, and M. Vingron. Identifying splits with clear separation: a new class discovery method for gene expression data. *Bioinformatics*, 17 Suppl 1:S107–14, 2001.

[25] W. Weaver and C. E. Shannon. *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, Illinois, 1949. Published in paperback 1963.

[26] M. Weber. Clustering by using a simplex structure. Technical report, ZR-04-03, Zuse Institute Berlin, February 2004.

[27] M. Weber and T. Galliat. Characterization of transition states in conformational dynamics using Fuzzy sets. Technical Report Report 02–12, Zuse Institute Berlin (ZIB), March 2002.

[28] E. Xing and R. Karp. CLIFF: clustering of high-dimensional microarray data via iterative feature filtering. *Bioinformatics*, 17:306–315, 2001.