Project acronym	SIMBAD
Project full title	Beyond Features: Similarity-Based Pattern Analysis and Recognition
Deliverable Responsible	Department of Computer Science ETH Zurich Universitätstrasse 6 http://www.ml.inf.ethz.ch/
Project web site	http://simbad-fp7.eu
EC project officer	Teresa De Martino
Document title	Report on Structure Preserving Embedding of Dissimilarity Data
Deliverable n.	D4.2
Document type	Report
Dissemination level	Public
Contractual date of delivery	M (month 24.)
Project reference number	213250
Status & version	Definitive version
Work package	WP4
Deliverable responsible (SHORT NAME)	ETHZ
Contributing Partners (SHORT NAME)	
Author(s)	Peter Schuffler, Sharon Wulff, Joachim M. Buhmann, Cheng Soon Ong, Volker Roth
Additional contributor(s)	

SIMBAD
Deliverable 4.2: Report on Structure Preserving Embedding of Dissimilarity Data

Peter Schüffler, Sharon Wulff,
Joachim M. Buhmann, Cheng Soon Ong, Volker Roth

Introduction

For several major applications in data mining, data is often not available as feature vectors in a vector space. For instance, genomics typically produce data represented as strings from some alphabet, psychology yields sets of similarity judgments, yet other fields like social sciences measure so called preference data. The missing vector space representation precludes the use of well established machine learning techniques such as Principal Component Analysis [1] or Support Vector Machines [2].

A common approach to handling Non-vectorial datasets is to replace the initial data by a collection of real numbers representing some "comparison" among the elements of the dataset. This procedure yields a matrix gathering the pairwise relations between the original objects, which may be the starting point of further data analysis.

The clustering approaches discussed in this report aim at identifying subsets or clusters of objects represented as "blocks" in a permuted dissimilarity matrix. The underlying idea is that objects grouped together in such a cluster can be reasonably well described as a homogeneous sub-population. Our focus on dissimilarity matrices implies that we do not have access to a vectorial representation of the objects, and in general, no such representation will exist, since we do not assume that the dissimilarity matrix fulfills the axioms of a valid metric.

In this report we summarize our studies on embedding strategies in the context of clustering. In the first part of this document we will mainly summarize our results for the pairwise k-means clustering cost function as outlined in [3]: we begin with a short overview of proximity-based data grouping, and then we focus on reformulating such problems with vectorial data representations. For the class of pairwise clustering methods that are related to minimizing a shift-invariant cost function, the *constant shift embedding* procedure is presented. A surprising property of this embedding is the complete preservation of group structure. The original non-metric pairwise clustering problem can be restated as a grouping problem over points in a vector space, yielding identical assignments of objects to clusters. Using the constant-shift embedding principle, we then demonstrate the equivalence between the *pairwise clustering* cost function and the classical k-means grouping criterion in the embedding space. The conclusion is

that the the k-means cost function (or its dissimilarity-based counterpart) is essentially "blind" against metric violations.

In the second part we will analyze a more general setting where the hard-clustering scenario with fixed number of clusters is replaced by a probabilistic approach which is capable of selecting the number of clusters in a data-adaptive way. We show that this probabilistic model is shift invariant only in an approximate sense, and in particular we show that exact shift invariance and data-adaptive selection of the number of clusters define two conflicting goals.

We conclude this report with a (sober) discussion about the role of structure preserving embeddings for the overall goal in the SIMBAD project, namely for building a novel theory for similarity-based pattern recognition.

1 Constant Shift Embedding for Pairwise Clustering

1.1 Proximity-based clustering

Unsupervised grouping or *clustering* aims at extracting hidden structure from data [4]. The term data refers to both a set of objects and a set of corresponding object representations resulting from some physical measurement process. Different types of object representations are possible, the two most common of which are *vectorial data* and *pairwise proximity data*. In the first case, a set of n objects is represented as n points in a d-dimensional vector space, whereas in the second case we are given a $n \times n$ pairwise proximity matrix.

The problem of grouping vectorial data has been widely studied in the literature, and many clustering algorithms have been proposed [4, 5]. One of the most popular method is k-means clustering. It derives a set of k prototype vectors which quantize the data set with minimal quantization error.

Partitioning proximity data is considered a much harder problem, since the inherent structure is hidden in n^2 pairwise relations. This datatype, however, is abundant in many applications, such as molecular biology, psychology, linguistics etc. In general, the proximities can violate the requirements of a distance measure, i.e. they may be non-symmetric and negative, and the triangle inequality does not necessarily hold. Thus, a loss-free embedding into a vector space is not possible, so that grouping problems of this kind can not directly be transformed into vectorial problems by means of classical embedding strategies.

Among several methods for clustering proximity-based data, in this first part of the document we will focus on those techniques that explicitly minimize a certain cost function. This subset of clustering methods includes e.g. graph-theoretic approaches like several variations of *Cut* criteria [6], and several methods derived from an axiomatization of pairwise cost functions in [7]. From a theoretical viewpoint, cost-based clustering methods are interesting insofar, as many properties of the grouping solutions can be derived by analyzing invariance properties of the cost function.

Among the class of cost-based criteria, the main focus of this work concerns those cost functions which are invariant under constant additive shifts of the pairwise dissimilarities. For this subset of clustering criteria we show that there always exists a set

of vectorial data representations such that the grouping problem can be equivalently restated in terms of Euclidean distances between these vectors. A special cost function of this kind is the *pairwise clustering cost function*. It is of particular interest, since it combines the properties of additivity, scale- and shift invariance, and statistical robustness, see [7]. In [8] this grouping problem is stated as a combinatorial optimization problem, which is optimized in a *deterministic annealing* framework after applying a mean-field approximation.

According to the theorem 2, we can always find a vectorial data representation such that the optimal partitioning w.r.t. the pairwise cost function is *identical* to k-means partitioning in the embedding space. This property demonstrates that the embedding method is optimal w.r.t. to distortions of the *data partition*. This distortion preserving embedding has to be contrasted with alternative, in our view not consistent, approaches that are optimal w.r.t. some *a priori* chosen MDS distortion measure.

Formulating pairwise clustering as a k-means problem yields several advantages, both of theoretical and technical nature: (i) the availability of prototype vectors defines a generic rule for using the learned partitioning in a predictive sense, (ii) we can apply standard noise- and dimensionality-reduction methods in order to separate the "signal" part of the data from underlying "noise", (iii) fast and efficient local search heuristics for optimizing the clustering cost functional often work much better in low dimensional embedding spaces.

1.2 The Pairwise Clustering Cost function

The modeling idea behind the Pairwise Clustering cost function is to minimize the sum of *pairwise* intra-cluster distances, emphasizing *compact* clusters. Optimizing a compactness criterion is certainly a very intuitive meta-principle for exploratory data analysis. It should be noticed, however, that other such meta-principles have been proposed, such as *separation* measures, mixed *compactness/separation* measures or *connectivity* measures. In order to formalize Pairwise Clustering, we define for each object a binary assignment variable that indicates its cluster membership. Let these variables be summarized in the $(n \times k)$ binary stochastic assignment matrix $M \in \{0,1\}^{n \times k} : \sum_{\nu=1}^{k} M_{i\nu} = 1$. Given a $(n \times n)$ dissimilarity matrix D, the Pairwise Clustering cost function reads:

$$H^{\text{pc}} = \frac{1}{2} \sum_{\nu=1}^{k} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} M_{i\nu} M_{j\nu} D_{ij}}{\sum_{l=1}^{n} M_{l\nu}}. \tag{1}$$

The optimal assignments $\hat{M}$ are obtained by minimizing H^{pc}. The minimization itself is a $\mathcal{NP}$ hard problem [9], and some approximation heuristics have been proposed: in [8] a *mean field annealing* framework has been presented. In [7] it has been shown that the time-honored *Ward's method* can be viewed as a hierarchical approximation of H^{pc}.

1.3 A special case: k-means clustering

For the special case of squared Euclidean distances between vectors $\{x_i\}_{i=1}^n, \quad x_i \in \mathbb{R}^d$, it is well known that H^{pc} is identical to the classical k-means cost function, see [4]. We now briefly review this relationship. The k-means cost function is defined as

$$H^{\text{km}} = \sum_{\nu=1}^{k} \sum_{i=1}^{n} M_{i\nu} \|x_i - y_\nu\|^2. \tag{2}$$

It measures the sum of squared intra-cluster distances to the prototype vectors

$$y_\nu := \frac{\sum_{i=1}^n M_{i\nu} x_i}{n_\nu}, \tag{3}$$

where $n_\nu := \sum_{l=1}^n M_{l\nu}$ denotes the number of objects in cluster ν. H^{km} can be written in a pairwise fashion by exploiting a simple algebraic identity for squared Euclidean distances:

$$\begin{aligned} \|x_i - y_\nu\|^2 &= \frac{1}{n_\nu} \sum_{j=1}^{n} M_{j\nu} \|x_i - x_j\|^2 - \frac{1}{2n_\nu^2} \sum_{j=1}^{n} \sum_{l=1}^{n} M_{j\nu} M_{l\nu} \|x_j - x_l\|^2, \\ \sum_{i=1}^{n} M_{i\nu} \|x_i - y_\nu\|^2 &= \frac{1}{2n_\nu} \sum_{j=1}^{n} \sum_{l=1}^{n} M_{j\nu} M_{l\nu} \|x_j - x_l\|^2. \end{aligned} \tag{4}$$

Substituting the latter identity into (2), we obtain

$$H^{\text{km}} = \frac{1}{2} \sum_{\nu=1}^{k} \frac{\sum_{i=1}^n \sum_{j=1}^n M_{i\nu} M_{j\nu} \|x_i - x_j\|^2}{\sum_{l=1}^n M_{l\nu}} = H^{\text{pc}}. \tag{5}$$

From this viewpoint, k-means clustering can be interpreted as a method for minimizing the sum of squared *pairwise* intra-cluster distances $D_{ij} = \|x_i - x_j\|^2$. The reader should notice, however, that in the general case of arbitrary dissimilarities D_{ij} a direct algebraic re-transformation of H^{pc} into H^{km} is *not* possible. Despite this fact, we will show in the remainder of this paper that it is still possible to obtain the optimal assignment variables $\hat{M}$ with respect to $H^{\text{pc}}(M)$ by minimizing a suitably transformed k-means problem. The key ingredient will be the *shift invariance property* of the Pairwise Clustering cost function: H^{pc} is invariant (up to a constant) under additive shifts of the *off-diagonal* elements of the dissimilarity matrix:

$$\tilde{D}_{ij} = D_{ij} + d_0(1 - \delta_{ij}) \quad \Rightarrow \quad \tilde{H} = H + (1/2) \cdot (n-k) d_0 = H + \text{const.} \tag{6}$$

Note that the optimal assignments of objects to clusters are not influenced by adding a constant to the cost function, i.e. $\hat{M}(\tilde{D}) = \hat{M}(D)$.

1.4 Constant shift embedding

We have introduced the cost function H^{pc} as a special instance of pairwise clustering problems. Due to the shift-invariance property (6), the partitioning of the dataset

(i.e. the assignments of a set of n objects to k clusters) is not affected by a constant additive shift on the off-diagonal elements of the pairwise dissimilarity matrix $D = (D_{ij}) \in \mathbb{R}^{n \times n}$. In the remainder of this paper, we will consider general dissimilarity matrices D, restricted only by the constraint that all self-dissimilarities are zero, i.e. that D has zero diagonal elements. We show that by exploiting the above shift invariance we can always embed such data into a Euclidean space without influencing the cluster structure. An off-diagonal shifted dissimilarity matrix reads

$$\tilde{D} = D + d_o\big(e_n e_n^t - I_n\big) \tag{7}$$

where $e_n = (1, 1, \dots 1)^t$ is a n-vector of ones and I_n the $n \times n$ identity matrix. In other words, (7) describes a constant additive shift $\tilde{D}_{ij} = D_{ij} + d_o \quad \text{for all } i \neq j$.

Before developing the main theory, we have to introduce the notion of a *centralized matrix*. Let P be an $(n \times n)$ matrix and let $Q = I_n - \frac{1}{n} e_n e_n^t$. Q is the projection matrix on the orthogonal complement of e_n. Define the *centralized* P by:

$$P^c = QPQ. \tag{8}$$

A centralized matrix has row- and column-sum equal to zero, which can easily be seen by looking at the components of P^c

$$P^c_{ij} = P_{ij} - \frac{1}{n}\sum_{k=1}^{n} P_{ik} - \frac{1}{n}\sum_{k=1}^{n} P_{kj} + \frac{1}{n^2}\sum_{k,l=1}^{n} P_{kl}. \tag{9}$$

Let us now consider symmetric dissimilarity matrices. Given such a symmetric and zero–diagonal matrix D, we decompose it the following way by introducing a new matrix S:

$$D_{ij} = S_{ii} + S_{jj} - 2S_{ij}. \tag{10}$$

It is clear that this decomposition is not unique unless we specify the diagonal elements of S. Let $\mathbb{S}_D$ denote the equivalence class of all S yielding the same D. The following lemma states that for all members $S \in \mathbb{S}_D$ the centralized version S^c is identical and uniquely defined by the given matrix D:

Lemma 1. *For any symmetric and zero–diagonal matrix D the following holds:*

$$S^c = -\frac{1}{2} D^c, \text{ with } D^c = QDQ.$$

The matrix S^c is a particularly interesting member of $\mathbb{S}_D$, since the following theorem holds:

Theorem 1. *D derives from a squared Euclidean distance, i.e. $D_{ij} = \|x_i - x_j\|^2$, if and only if S^c is positive semi-definite.*

Proof. [10] referring to [11]. □

For general dissimilarities, S^c will be indefinite. By shifting its diagonal elements, however, we can transform it into a positive semi-definite matrix: the following lemma states that for any matrix A, a positive semi-definite matrix $\tilde{A}$ can be derived by subtracting the smallest eigenvalue from all of its diagonal elements:

Lemma 2. *Let $\tilde{A} = A - \lambda_n(A)I_n$, where $\lambda_n(\cdot)$ is the minimal eigenvalue of its argument. Then $\tilde{A}$ is positive semi-definite.*

Proof. Due to the diagonal shift, the smallest eigenvalue becomes zero. □

We can now summarize the above results: given a matrix D, there exists a unique matrix S^c by lemma 1. If S^c is not positive semi-definite, lemma 2 states that by subtracting $\lambda_n(S^c)$ from its diagonal elements, we obtain a positive semi-definite $\tilde{S}$. Returning to (10) with our fixed matrix S^c, such a diagonal shift of S^c corresponds to an *off-diagonal* shift of the dissimilarities

$$\tilde{D}_{ij} = \tilde{S}_{ii} + \tilde{S}_{jj} - 2\tilde{S}_{ij} \;\Leftrightarrow\; \tilde{D} = D - 2\lambda_n(S^c)\big(e_n e_n^t - I_n\big). \tag{11}$$

In other words, if we were given $\tilde{D}$ instead of our original D, then $\tilde{S}$ would be a positive semi-definite member of the equivalence class $\mathbb{S}_{\tilde{D}}$ of matrices fulfilling the decomposition $\tilde{D}_{ij} = \tilde{S}_{ii} + \tilde{S}_{jj} - 2\tilde{S}_{ij}$. Theorem 1 then tells us that this off-diagonally shifted matrix $\tilde{D}$ derives from a squared Euclidean distance. Since every positive semi-definite matrix is a dot product– (or *gram*–) matrix in some vector space, there exists a matrix X of vectors such that $\tilde{S} = XX^t$. The matrix $\tilde{D}$ then contains squared Euclidean distances between these vectors. We can now insert $\tilde{D}$ into our clustering procedure (which is assumed shift-invariant), and we will obtain the same partition of the objects as if we had clustered the original matrix D. Contrary to directly using D, however, the matrix $\tilde{D}$ now contains squared Euclidean distances between a set of vectors $\{x_i\}_{i=1}^n$. The vectors themselves can be reconstructed by way of kernel PCA, see [12].

A k-means formulation for Pairwise Clustering It is well-known that for the special case of squared Euclidean distances, the Pairwise cost function and the k-means cost function can be transformed into each other by using a simple algebraic identity, see above. The invariance property in eq. (6), however, implies that a similar relationship between both cost functions holds in the general setting:

Theorem 2. *Given an arbitrary $(n \times n)$ dissimilarity matrix D with zero self-dissimilarities, there exists a transformed matrix $\tilde{D}$ such that*
(i) the matrix $\tilde{D}$ can be interpreted as a matrix of squared Euclidian distances between a set of vectors $\{x_i\}_{i=1}^n$ with dimensionality dim$(x_i) \leqslant n - 1$,
(ii) the original pairwise clustering problem defined by the cost function $H^{pc}(D)$ is equivalent to the k-means problem with cost function H^{km} in this vector space, i.e. the optimal cluster assignment variables $\hat{M}_{i\nu}$ are identical in both problems: $\hat{M}^{pc}(D) = \hat{M}^{km}(\tilde{D})$.

2 A Probabilistic Generalization: the Wishart-Dirichlet Cluster Process

Despite its elegance, the approach described above is particularly tailored to certain hard-clustering cost functions like the pairwise k-means function. Here we go one step further and reformulate the matrix partitioning problem in a fully probabilistic framework. Clustering with such models can be viewed as a low-rank matrix approximation, and approximate shift invariance can be explained as a natural consequence of assuming a white noise term capturing the deviations from the low-rank model. In the hard-clustering limit, the k-means model with its known invariance properties appears as a special case of this class of models.

This section is structured as follows: we first review the partitioning model for Gaussian mixtures introduced in [13], which is then extended to a partitioning process on matrices. Connections to multi-dimensional scaling are shown which help to explain the clustering process as a low-rank matrix approximation. Finally, shift invariance properties are analyzed, and the model is tested both on synthetic and real-world data.

2.1 Gauss-Dirichlet cluster process

Let $[n] := \{1, \ldots, n\}$ denote an index set, and $\mathbb{B}_n$ the set of partitions of $[n]$. A partition $B \in \mathbb{B}_n$ is an equivalence relation $B : [n] \times [n] \to \{0, 1\}$ that may be represented in matrix form as $B(i, j) = 1$ if $x(i) = x(j)$ and $B(i, j) = 0$ otherwise, with x being a function that maps $[n]$ to some label set $\mathbb{L}$. Alternatively, B may be represented as a set of disjoint non-empty subsets called "blocks" b. A *partition process* is a series of distributions P_n on the set $\mathbb{B}_n$ in which P_n is the marginal distribution of P_{n+1}. Such a process is called *exchangeable* if each P_n is invariant under permutations of object indices, see [14] for details.

A *Gauss-Dirichlet cluster process* consists of an infinite sequence of points in $\mathbb{R}^d$, together with a random partition of integers into k blocks. A sequence of length n can be sampled as follows, cf. [13, 15, 16]: fix the number of mixture modes k, generate mixing proportions $\pi = (\pi_1, \ldots, \pi_k)$ from an exchangeable Dirichlet distribution $\mathrm{Dir}(\lambda/k, \ldots, \lambda/k)$, generate a label sequence $(x_1, \ldots, x_n)$ from a multinomial distribution, and forget the labels introducing the random partition B of $[n]$ induced by x. Integrating out π, one arrives at a Dirichlet-Multinomial-type prior over partitions:

$$P_n(B|\lambda, k) = \frac{k!}{(k - k_B)!} \frac{\Gamma(\lambda) \prod_{b \in B} \Gamma(n_b + \lambda/k)}{\Gamma(n + \lambda)[\Gamma(\lambda/k)]^{k_B}}, \tag{12}$$

where $k_B \leq k$ denotes the number of blocks present in the partition B and n_b is the size of block b. The limit as $k \to \infty$ is well defined and known as the Ewens process (a.k.a. Chinese Restaurant process), see for instance [17, 18, 19]. Given such a partition B, d-dimensional observations $Y = (Y_1, \ldots, Y_n)$ are generated from a zero-mean Gaussian distribution with covariance matrix

$$\Sigma_B = I_n \otimes \Sigma_0 + B \otimes \Sigma_1, \quad \text{with} \quad \mathrm{cov}(\mathcal{Y}_{ir}, \mathcal{Y}_{js}|B) = \delta_{ij}\Sigma_{0rs} + B_{ij}\Sigma_{1rs}, \tag{13}$$

where Σ_0 is the usual within-class covariance matrix and Σ_1 the between-class matrix, respectively. Since the partition process is invariant under permutations, we can always think of B being block-diagonal. For spherical covariance matrices, $\Sigma_0 = \alpha I_d, \Sigma_1 = \beta I_d$, the columns of $\mathcal{Y}$ contain independent copies distributed according to a normal distribution with covariance matrix $\Sigma_B = \alpha I + \beta B$. Further, the distribution also factorizes over the blocks $b \in B$. Introducing for each block a $(n_b \times n_b)$-matrix of ones E_{n_b}, the joint distribution of data and partitions reads

$$p(Y, B|\alpha, \beta, \lambda, k) = \left[\prod_{b\in B} \prod_{j=1}^{d} N(Y_{i_b j}|\alpha I_{n_b} + \beta E_{n_b})\right] \cdot P(B|\lambda, k), \qquad (14)$$

where the symbol i_b defines an index-vector for all objects assigned to block b.

2.2 Wishart-Dirichlet cluster process

We now extend the Gauss-Dirichlet cluster process to a sequence of inner-product and distance matrices. Assume that the random matrix $\mathcal{Y}_{n\times d}$ follows the zero-mean Gaussian distribution specified in (13), with $\Sigma_0 = \alpha I_d, \Sigma_1 = \beta I_d$. Then, conditioned on the partition B, the inner product matrix $\mathcal{S} = \mathcal{Y}\mathcal{Y}^t/d$ follows a (possibly singular) Wishart distribution in d degrees of freedom, $\mathcal{S} \sim \mathcal{W}_d(\Sigma_B)$, [20]. If we directly observe S (i.e. if we measure similarities expressed as a Mercer kernel matrix), it suffices to consider the conditional probability of partitions, $P_n(B|S)$, which has the same functional form for ordinary and singular Wishart distributions. Due to the block structure in B, $P_n(B|S)$ factorizes over the blocks $b \in B$:

$$P_n(B|S, \alpha, \beta, \lambda, k) \propto \left[\prod_{b\in B} |\Sigma_b|^{-\frac{d}{2}} \exp\left(-\tfrac{d}{2}\mathrm{tr}(\Sigma_b^{-1} S_b)\right)\right] \cdot P_n(B|\lambda, k), \qquad (15)$$

where Σ_b, S_b denote the submatrices corresponding to the b-th block.

Often, however, we do not directly observe S, but only a matrix D of squared distances with components $D_{ij} = S_{ii} + S_{jj} - 2S_{ij}$. Note that S determines D, but not vice versa, since D is constant on equivalence classes of S resulting from the arbitrariness of the mean vector. A squared distance matrix D is characterized by the property of being *negative definite on contrasts*, which means that $\boldsymbol{x}^t D\boldsymbol{x} = -\frac{1}{2}\boldsymbol{x}^t S\boldsymbol{x} < 0$ for any $\boldsymbol{x} : \boldsymbol{x}^t\mathbf{1} = 0$. The distribution of D has been formally studied in [21], where it was shown that if $\mathcal{S} \sim \mathcal{W}_d(\Sigma_B)$, $-D$ follows a generalized Wishart distribution, $-\mathcal{D} \sim \mathcal{W}_d(\mathbf{1}, \Delta)$ defined with respect to the transformation kernel $\mathbf{1}$, where $\Delta_{ij} = \Sigma_{Bii} + \Sigma_{Bjj} - 2\Sigma_{Bij}$. As before, the transformation kernel has the effect that the distribution of D is constant on equivalence classes. Since we are interested in studying the partition B given an observed matrix D, it is convenient to forego the equivalence classes by explicitly choosing a representation S which fulfills $D_{ij} = S_{ii} + S_{jj} - 2S_{ij}$. We can again use the projection Q with $Q_{ij} = \delta_{ij} - \frac{1}{n}$ to transform D into centered inner product form via $S = -\frac{1}{2}QDQ$, which eliminates contributions of the mean vector while preserving the distances. Formally, this choice is justified by the observation that D is a matrix of squared distances if and only if $S = -\frac{1}{2}QDQ$ is positive semi-definite [10].

Relation to multi-dimensional scaling. Classical multi-dimensional scaling [22] can be interpreted as using a distance model

$$-D \sim \mathcal{W}(\mathbf{1}, \Delta) \text{ with } \Delta = \Delta_0 - M - \sigma^2 I, \tag{16}$$

where Δ_0 stems from the transformation kernel $\mathbf{1}$, M is a low-rank matrix used to approximate the observed matrix D, and $\sigma^2 I$ is a white noise term accounting for deviations from the low-rank model, see [21]. As before, the transformation $S = -QDQ$ eliminates the contribution of the kernel and transforms the data into inner product form. The matrix M is then computed as the best low-rank approximation to S, which is the rank-constrained maximum likelihood solution in the Wishart model, see [21]. The above expression $\Sigma = \sigma^2 I + M$ is essentially the same as our covariance model $\Sigma_B = \alpha I + \beta B$. The only difference is that B is not an arbitrary low-rank matrix, but additionally constrained to be a binary partition matrix. Thus, our partitioning model can be understood as a binarized version of multi-dimensional scaling. The white noise term αI corresponding to the within-class covariance has the role of absorbing the deviations from the low-rank model.

Shift invariance. The expected value of $\mathcal{S} \sim \mathcal{W}_d(\Sigma_B)$ is $E[\mathcal{S}] = \Sigma_B$. Adding an additional noise term δI shifts the expected value to $\Sigma_B + \delta I$. Reversing this argument for inference problems in which we observe the inner product matrix S, additive shifts of the diagonal elements of S might be absorbed by the white noise term. Note that such additive diagonal terms appear when shifting the *off*-diagonal elements of D. Using sufficiently large shifts ensures that there exists an embedding space in which the transformed dissimilarities D' can be represented as squared Euclidean distances. The idea behind additive shifts is the following: if we observe a matrix D which gives rise to an indefinite matrix $S = -\frac{1}{2}QDQ$, there are basically two options: either we can directly use S, irrespective of negative eigenvalues, or we can try to "heal" the negative eigenvalues. Concerning the first option, it is unclear what bias is introduced due to the model mismatch. "Healing" the negative eigenvalues, on the other hand, introduces another sort of bias. In the ideal case, we can find a transformation which exploits some invariance of the analysis model. If the model is invariant under additive shifts, we can safely transform *any* (symmetric) matrix D in such a way that it will be inside the model space. Note that for our clustering model, even symmetry is not required, since all conditionals are invariant under $S \leftarrow 1/2(S + S^t)$. We first show that exact shift invariance is possible, but only under assumptions that eliminate the probabilistic nature of the model.

The inverse matrix $\Sigma_b^{-1} = (\alpha I_{n_b} + \beta E_{n_b})^{-1}$ can be analytically computed as

$$\tfrac{1}{\alpha}\left(I - \tfrac{\beta}{\alpha + n_b\beta} E_{n_b}\right) = \tfrac{1}{\alpha}\left(I - \tfrac{\theta}{1 + n_b\theta} E_{n_b}\right) \quad \text{with } \theta := \beta/\alpha. \tag{17}$$

Denoting by $\frac{d}{2}A$ the argument of the exponential function in (15), a shift $S' = S + \delta I_n$ implies

$$\alpha A' := -\alpha \operatorname{tr}((\alpha I_{n_b} + \beta E_{n_b})^{-1}(S_b + \delta I)) = \tfrac{n_b\theta}{1+n_b\theta}(n_b\bar{S}_b + \delta) - \operatorname{tr}(S_b) - n_b\delta, \tag{18}$$

where $\bar{S}_b$ denotes the mean value of the b-th block of S. For $\alpha \to 0$, it follows that

$$\alpha A' \approx n_b \bar{S}_{n_b} - \mathrm{tr}(S_b) - (n_b - 1)\delta \Rightarrow \sum_{b \in B} \alpha A' \approx -\mathrm{tr}(S) - (n - k_B)\delta + \sum_{b \in B} n_b \bar{S}_b \tag{19}$$

with k_B being the number of blocks in the partition B. This result implies that for fixed α, θ, k_B, the conditional posterior of partitions is approximately shift invariant. In the hard-clustering limit as $\alpha \to 0$ this statement becomes exact. The price for exact shift invariance is the problem of estimating k_B. The restriction to hard assignments precludes an intrinsic measure of "clusterability": the model degenerates to an combinatorial optimization problem in which we need to fix k. The optimal solution will then automatically include all $k_B = k$ blocks. Note that the limit $\alpha \to 0$ defines the *pairwise clustering* cost function [8] whose invariance properties have been studied in [3].

Here, we consider more realistic situations in which both the covariance parameters and k_B are estimated. Intuitively, we assume that shifts are "absorbed" in the within-class term, i.e. $\alpha' = \alpha + \delta$. Analytically studying the effects on the partition when both α and θ are varying is complicated, in particular due to the influence of the normalization term $|\Sigma_b|^{-(d/2)}$ in (15). Thus, we only consider an idealized scenario in which the matrix S has a distinct cluster structure which is consistent with our model. In such a case, there will be a matrix $\Sigma' = \alpha' I + \beta B'$ that is reasonably close to the observed S, and the ML-estimate of the covariance matrix in the Wishart model is $\hat{\Sigma}_B \approx \Sigma' = \alpha' I + \beta B'$. If there is an additional shift $S^{\text{shifted}} = S + \delta I$, the ML-estimate will be $\hat{\Sigma}_B^{\text{shifted}} \approx (\alpha' + \delta) I + \beta B'$. The normalization term, however, decreases, indicating that the distribution is smeared out due to the increased noise term. Note that we have neglected the influence of the prior $P_n(B)$ defined in eq. (12). For moderate shifts, however, the deviations from "local" uniformity might be reasonably small. Despite the approximate nature of this plausibility argument, our simulation experiments nicely corroborate the intuition that moderate shifts can be absorbed in the white-noise term — at least if the data exhibits a clear cluster structure. In practice, however, observed matrices only rarely show a distinct block structure, and the additional noise component introduced by large shifts severely hampers the estimation of a stable partition, both for our probabilistic model and for the hard-clustering counterpart. Thus, the real benefit of any form of shift invariance might be a justification for first transforming the data into inner product form and then applying (kernel-)PCA-denoising to eliminate the additional noise, which is exactly the approach suggested in [23].

Inference via Gibbs sampling. The main idea in Gibbs sampling is to iteratively sample parameter values from the full conditionals, For the sake of simplicity, we only consider the update equations for the partition B. Assume that n objects in S have already been partitioned according to B. Conditioning on S and B, we want to compute the assignment probabilities for a *new* object o_*, characterized by an additional row and column in the augmented matrix S_*. Due to permutation invariance, we can always assume that S_* is ordered according to blocks in B and that the additional row/column is the last one in some block. Either the new object is assigned to an existing block b, i.e. $o_* \to b \in B$, or it is assigned to a new block which will be denoted by $o_* \to \emptyset$.

Consider first the case $o_* \to b \in B$. Assume that the new row/column is the last

one in this block. The number of objects in block b is increased by one, i.e. $n_b^* = n_b+1$, and the new block mean is denoted by $\bar{S}_b^*$. With a slight abuse of notation, we write S_{*j} for $S^*_{n_b+1,j}$ and S_{**} for $S^*_{n_b+1,n_b+1}$. All symbols without $(*)$ refer to the old state with n objects. Denote by $\frac{d}{2}A^*(b)$ the new argument in the exponential function in (15). Then,

$$A^*(b) = A + \tfrac{1}{\alpha}\left(\tfrac{(n_b+1)\theta}{1+(n_b+1)\theta}(n_b+1)\bar{S}_b^* - \tfrac{n_b\theta}{1+n_b\theta}n_b\bar{S}_b + S_{**}\right). \tag{20}$$

Consider now the case of assigning o_* to a new cluster, i.e. $o_* \to \emptyset$. A new singleton cluster is added, i.e. $k_B^* = k_B+1$. The associated argument in the exponential function becomes: $A^*(\emptyset) = A + \frac{1}{\alpha}\left(\frac{\theta}{1+\theta}S_{**} + S_{**}\right)$. For the conditionals, we need to multiply the exponentiated terms above with the contributions of both the normalization term in (15) and the prior. Denoting these terms by $N^*(b)$ and $N^*(\emptyset)$, and using $\Gamma(x+1) = x\Gamma(x)$ in (12), we find

$$N^*(b) \propto \left[\tfrac{1+\theta n_b}{1+\theta(1+n_b)}\right]^{(d/2)} \cdot (n_b + \lambda/k), \quad N^*(\emptyset) \propto (1+\theta)^{-\frac{d}{2}} \cdot \lambda(1-k_B/k). \tag{21}$$

2.3 Experiments

In a first experiment we analyze the shift-invariance based on a matrix sampled from $\mathcal{W}(\Sigma_B)$ with a two-block partition (30%/70%) and $\alpha = 1, \theta = 20$. Using relatively uninformative priors on α and θ, we add increasing shifts δI to S. To compensate for δ, we adjust the priors over α and θ by shifting their expected value accordingly. Figure 1 shows that over a large range of δ-values, the shift is indeed absorbed in α, and the estimate for $\beta = \alpha \cdot \theta$ is roughly constant. Deviations from the "true" partition are summarized in the expression $\sum_{ij}(B_{ij}^{\text{true}} - B_{ij}^{\text{sampled}})$. Note that even for large shifts ($\delta = 1000$ is roughly 25% of the largest eigenvalue of S), the partition remains rather stable. It is clear that we consider an idealized scenario, but nevertheless we conclude that our intuition about absorbing shifts seems to be correct. In this experiment the influence of λ is extremely small: λ can be changed over at least 10 orders of magnitude without affecting k_B.

In the two following experiments we quantitatively investigate the clustering performance in terms of the size-normalized within-sum-of-squared errors (distances), $SSE = \sum_{b\in B} n_b \bar{D}_b$, and compare the outcome with the Affinity Propagation (AP) method based on two datasets described in [24]. The first dataset contains similarities between 900 face images from the Olivetti database, available at *http://www.psi.toronto.edu/affinitypropagation/*). AP has been reported to exhibit some advantages over other centroid-based approaches on this dataset. The results in Figure 2 (top row) show that our model consistently outperforms AP, which means that better centroids have been found (note, however, that in this comparison our model has the advantage of not being restricted to choosing exemplars as centroids). Nevertheless, we conclude that in terms of optimization quality, the Wishart-Dirichlet model is a strong competitor to AP. Even more important, however, is the observation that all partitions with $k_B < 100$ are very implausible, because such a low number of clusters can only be obtained by "forcing" the model to use a very low θ-value via the prior,

see the right panel: sampled values "hitting" the upper boundary of admissible values indicate that the model is entirely forced into a certain direction by the prior. Note that θ is the quotient of between-class to within-class variance, and $\theta < 1$ means that there is hardly any cluster structure in the data.

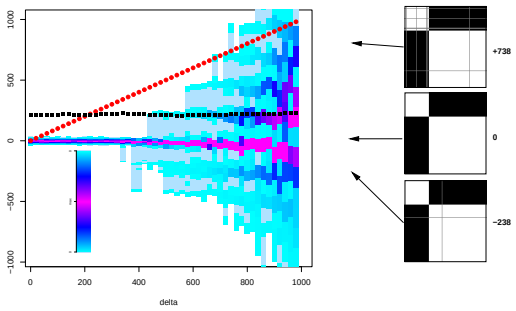

Figure 1: Toy example for analyzing shift invariance. Linearly increasing red circles: mean values of estimated α-parameter under variation of the shift δ. Almost horizontal black squares: mean values of estimated β-parameter (scaled by a factor of 10 for better visualization). Color-coded histogram: differences between true and sampled partition. Right panel: three sampled partitions.

Using the AP model, on the other hand, we can simply "slide" through all k_B-values by changing the "affinity"-parameter from -74 to -15. From the AP model alone, we find it difficult to see why one of these results should be preferred over any other one (in [24] the model with $k_B = 62$ has been chosen for further analysis). The computational workload is not really an issue in this example, since even several millions of Gibbs sweeps can be computed reasonably fast (i.e. over night). A similar situation occurs for another dataset containing KL-divergences between sentences in a manuscript, which was used in [24] to demonstrate the performance of AP in situation where metric axioms are violated. Figure 2 (bottom row) clearly shows that (after symmetrizing an shifting) our model is a strong competitor in terms of optimization quality. The right panel again indicates that models with a low number of clusters (say < 70) are not very plausible due very small θ-values.

3 Conclusion

A partitioning model is called shift invariant, if the choice of a partition is not influenced by additive constant shifts of the off-diagonal elements in D. If a model exhibits this invariance property, it is always possible to construct an underlying Euclidean embedding space without altering the partition, a situation which we describe as "structure preserving embedding".

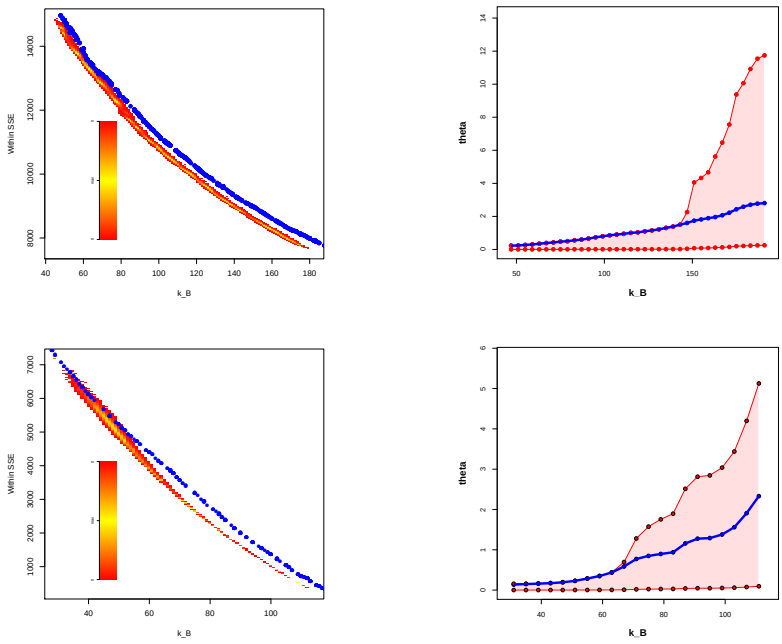

Figure 2: Clustering the face dataset (**top row**) and text dataset (**bottom row**) from [24]. **Left:** within SSE obtained from Affinity Propagation with different affinity-parameter values (blue circles) and from our algorithm under variation of the prior on θ (color-coded histogram). **Right:** Sampled θ values (blue) and range of possible θ values in the discretized prior (reddish-shaded area).

We have shown that the pairwise k-means cost function exhibits strict shift invariance, which –in terms of group structure– defines a structure preserving embedding model. However, this analysis is restricted to a certain cost function, and in particular to considering scenarios in which the number of clusters k is defined in advance. The latter requirement must be considered a severe shortcoming in most real applications, because information about the number of clusters usually rare. Therefore, we tried to broaden our viewpoint on pairwise clustering by considering a probabilistic version of the pairwise k-means model. The main idea is to construct a stochastic process on similarity matrices ad use a Dirichlet process prior to estimate the number of "blocks" in a partition matrix. Concerning structure preserving embeddings defined by constant-shift embeddings, we have shown that the clustering model induced by this Wishart-Dirichlet model can absorb "moderate" shifts in the white-noise term. However, a particular problem of this model is that the process of estimating the number of clusters in a data-adaptive fashion is also affected by the shift: shifting increases the tendency to introduce new clusters, since under the shift the mutual similarities between all objects decrease. It seems that strict shift invariance can only be achieved if the number of clusters is fixed, which somehow contradicts our efforts to generalize the k-means setting.

Considering the relevance of structure preserving embedding for the overall goal of the SIMBAD project, namely the development of a new theory of similarity-based pattern recognition, our current view is ambivalent: strict structure preservation could be proved only for a small set of clustering methods, like pairwise k-means and certain graph-based cut/association algorithms. All these algorithms require the user to fix the number of clusters in advance. A "relaxed" version of shift invariance holds for a probabilistic version of the pairwise k-means method, but we have to admit that shift invariance and estimation of the number of clusters might be two conflicting goals. As an alternative the number of clusters k can be estimated by the information theoretic approach to cluster valudation using approximation set coding (ASC) [25].

When it comes to building a theory on similarity-based pattern recognition, all these algorithms may be seen as "negative results", since they are essentially blind against Euclidean- or even metric violations. In other words: if one wants to learn something about clustering similarity data, one should look at different clustering procedures. While this result may be considered an interesting insight, it is still a very limited result due to the small number of algorithms that could be identified to fall into this category.

References

[1] I.T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1986.

[2] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2):181–201, 2001.

[3] V. Roth, J. Laub, M. Kawanabe, and J.M. Buhmann. Optimal cluster preserving embedding of non-metric proximity data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12), 2003.

[4] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern classification*. John Wiley & Sons, second edition, 2001.

[5] A.K. Jain, M.N. Murty, and P.J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[6] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[7] J. Puzicha, T. Hofmann, and J. Buhmann. A theory of proximity based clustering: Structure detection by optimization. *Pattern Recognition*, 33(4):617–634, 1999.

[8] T. Hofmann and J. Buhmann. Pairwise data clustering by deterministic annealing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(1):1–14, 1997.

[9] P. Brucker. On the complexity of clustering problems. In M. Beckman and H.P. Kunzi, editors, *Optimization and Operations Research: Lecture Notes in Economics and Mathematical Systems*, pages 45–54. Springer, 1978.

[10] W.S. Torgerson. *Theory and Methods of Scaling*. John Wiley and Sons, New York, 1958.

[11] G. Young and A. S. Householder. Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3:19–22, 1938.

[12] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

[13] P. McCullagh and J. Yang. How many clusters? *Bayesian Analysis*, 3, 2008.

[14] J. Pitman. Combinatorial stochastic processes. In J. Picard, editor, *Ecole d'Ete de Probabilites de Saint-Flour XXXII-2002*. Springer, 2006.

[15] S.N. MacEachern. Estimating normal means with a conjugate-style Dirichlet process prior. *Communication in Statistics: Simulation and Computation*, 23:727–741, 1994.

[16] D.B. Dahl. Sequentially-allocated merge-split sampler for conjugate and non-conjugate Dirichlet process mixture models. Technical report, Department of Statistics, Texas A&M University, 2005.

[17] W.J. Ewens. The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, 3, 1972.

[18] R.M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265, 2000.

[19] D. Blei and M. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1:121–144, 2006.

[20] M.S. Srivastava. Singular Wishart and multivariate beta distributions. *Annals of Statistics*, 2003.

[21] P. McCullagh. Marginal likelihood for distance matrices. *Statistica Sinica*, 19, 2009.

[22] T.F. Cox and M.A.A. Cox. *Multidimensional Scaling*. Chapman & Hall, 2001.

[23] V. Roth, J. Laub, J.M. Buhmann, and K.-R. Müller. Going metric: Denoising pairwise data. In S. Thrun S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 817–824. MIT Press, Cambridge, MA, 2003.

[24] Brendan J. Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.

[25] Joachim M. Buhmann. Information theoretic model validation for clustering. In *International Symposium on Information Theory, Austin Texas*. IEEE, 2010. (in press).