Project acronym	SIMBAD
Project full title	Beyond Features: Similarity-Based Pattern Analysis and Recognition
Deliverable Responsible	Prof. Dr. Joachim M. Buhmann Department of Computer Science, ETH Zurich, 8092 Zurich, Switzerland http://www.ml.inf.ethz.ch/
Project web site	http://simbad-fp7.eu
EC project officer	Teresa De Martino
Document title	Characterization of invariances
Deliverable	D 3.2
Document type	Report
Dissemination level	Public
Contractual date of delivery	M 12
Project reference number	213250
Status & version	Definitive version
Work package, deliverable responsible	WP 3, ETH Zurich
Author(s)	Peter Schüffler, Sharon Wulff, Joachim M. Buhmann, Cheng Soon Ong, Volker Roth
Additional contributor(s)	-

SIMBAD Deliverable D3.2

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

June 4, 2009

1 Similarity based pattern recognition and model selection

The search for patterns in data requires a mathematical definition of structure and a comparison function to rank different structures. This definition of structure would provide insights into the invariances in the problem class that we are considering. The central question raised in this work package WP3.2 is related to the information content of metricity violations in data. An information theoretic approach based on approximations is sketched in Section 2, the shift invariance in connection with a non-parametric Bayesian approach to pairwise clustering is studied in Section 3.

Section 2 summarizes an information theoretic view of cost approximation (empirical risk approximation, ERA) [1] and how a hypothetical communication framework can be used for model selection. It is well known that stability based model selection [4] yields highly satisfactory results in applications and a theoretical understanding of such a heuristics is highly desirable. Similarity based reasoning about structure in data is motivated by the uncertainty in data which induce uncertainty in the solution space. Patterns are considered to be similar if they are statistically indistinguishable due to data noise. The request that solutions should generalize from one data set to an equally probable second data set gives rise to a new notion of structure induced information. We use the problem of data clustering to illustrate the concept.

Despite the lack of a vectorial representation, one way to formulate clustering problems is to define a generative process that produces points in some Euclidean space according to some form of mixture distribution, and then study the distribution of the pairwise distance matrices associated with sets of points. In Section 3, we describe an approach that exploits inherent invariance principles of clustering models, by considering a matrix partitioning problem. This allows us to relax the metricity requirements. This fully probabilistic framework is used to show that an important class of mixture models can be viewed as low-rank matrix approximations, 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.

2 Tradeoff between informativeness and robustness

Structures are usually defined in statistical learning as elements of a hypothesis class, e.g., partitions in the case of clustering problems or projection matrices in case of linear dimension reduction. Depending on the application users then define a cost, risk or energy function to code the quality of such structures. The search for "good" or even optimal structures is then organized as an optimization procedure to reduce costs.

Such a modelling strategy does neither answer the open question which cost function should we use, nor does it help us to develop robust algorithms for finding such structures.

In the following, we describe an information theoretic view of cost approximation (empirical risk approximation, ERA) and how a hypothetical communication framework can be used for model selection. We use the problem of data clustering to illustrate the concept.

2.1 Statistical learning of clustering

Given are a **set of objects** $\mathcal{O} = \{o_1, \ldots, o_n\}$ and measurements $\mathbf{X} \in \mathcal{X}$ to characterize these objects. $\mathcal{X}$ denotes the measurement space. Such measurements might be vectors $\mathbf{x}_i \in \mathbb{R}^d, 1 \leq i \leq n$ in a d-dimensional space or relations $\mathbf{D} = (D_{ij}) \in \mathbf{R}^{n \cdot n}$ which describe the (dis)-similarity between object o_i and o_j. More complicated data structures than vectors or relations, e.g., three-way data or graphs, are used in various applications. In the following, we use the generic notation $\mathbf{X}$ for measurements.

The **hypothesis class** for a clustering problem is defined by the set of assignments of data to groups, i.e., $\mathcal{C} = \{c : \mathcal{O} \rightarrow \{1, \ldots, k\}^{|\mathcal{O}|}\}$. $\mathcal{C}(\mathbf{X})$ is a set of functions which map objects to cluster indices. For n objects we can distinguish $O(k^n)$ such functions. Special clustering models might require additional parameters θ which characterize a cluster like the centroids in k-means clustering. The hypothesis class is then the product space of possible assignments and possible parameter values.

Pattern analysis in data clustering requires to quantify the quality of such hypothesis, e.g., in vector quantization we use the k-means cost function and we use the nearest centroid assignment rule. For the subsequent discussion on empirical risk approximation we assume that a cost function $R(c, \theta; \mathbf{X})$ is given which measures how well a particular clustering with assignments $c(o)$ and cluster parameters θ groups the objects. A suitable metric on the space of hypotheses might be chosen based on such a cost function R.

2.2 Why information theory for clustering?

To formulate the statistical learning question we have to consider the following problem: Quite often the measurement space $\mathcal{X}$ has a much higher "dimension" than the solution space. Consider for example the problem of spectral clustering with k groups based on dissimilarities $\mathbf{D}$: The measurements are elements of $\mathbb{R}^{n(n-1)/2}$ for real valued weights, but we can only distinguish less than or equal to k^n different clusterings. Any approach which relies on estimating the probability

distribution of the data ultimately will fail since we require far too many observations than needed to identify one hypothesis or a set of hypotheses, i.e., one clustering or a set of clusterings.

Using an information theoretic perspective, we might ask the question how the uncertainty in the observations limit the resolution in the hypothesis class. How different can two hypotheses be so that they are still statistically indistinguishable given a cost function $R(c, \theta; \mathbf{X})$? The core question of statistical learning, "How well does a learning solution generalize?", is intimately related to the problem of distinguishing hypotheses.

Shannon's information theory provides a framework to study such questions of how many bit strings can be reliably distinguished in the presence of noise and, therefore, can be used as a code for communication. This study is based on the idea that approximation sets of clustering cost functions can be used as a reliable code. The capacity of such a coding scheme then answers the question how sensitive a particular cost function is to noise. "Good" models exhibit high robustness to noise and at the same time, they are highly informative due to a large hypothesis class. "Poor" models might be sensitive to noise (overfitting) or might be very restrictive with a small hypothesis class (underfitting).

To identify the correlate of a code vector and a code book, we define the set $\mathcal{C}_\gamma$ of hypotheses which are γ-optimal w.r.t. the minimum cost solution $c^\perp(\mathbf{X}) = \arg\min_{c,\theta} R(c, \theta; \mathbf{X})$, i.e.,

$$\mathcal{C}_\gamma(\mathbf{X}) = \{c \ : \ R(c, \theta; \mathbf{X}) \leq R(c^\perp, \theta^\perp; \mathbf{X}) + \gamma\}. \tag{1}$$

To test how well a γ-optimal solution generalizes we assume that we are given two samples of the problem, i.e., we have measurements $\mathbf{X}^{(1)}, \mathbf{X}^{(2)} \sim \Pr(\mathbf{X})$ for training and testing. These two measurements $X^{(1,2)}$ define two optimization problems $R(c, \theta; \mathbf{X}^{(\mathbf{1},\mathbf{2})})$. For both sets of measurements we can determine γ-optimal approximations $\mathcal{C}_\gamma^{(1,2)} = \mathcal{C}_\gamma(\mathbf{X}^{(1,2)})$. To measure stability of a solution, we require that the intersection between both sets is almost as large as its union. If this condition is met then the noise in the data will not affect the property to be γ-optimal to the optimum. γ-optimality can be considered as a similarity criterion based on the cost function $R(c, \theta; \mathbf{X})$ which closely relates this concept to the main theme of the `SIMBAD` project.

2.3 Coding by approximation

The informativeness-robustness tradeoff is expressed by the condition that γ should be as small as possible and the intersection between the two approximation sets should be as large as their union. This condition corresponds to Shannon's random coding argument that the received bit string should be jointly typical with the codeword which has been selected by the sender. The error of this communication process vanishes for asymptotically large bit strings provided we do not exceed the capacity of the communication channel.

In our setting, where we use approximation sets for coding, we have to generate $2^{n\rho}$ different code problems with respective approximation sets so that a zero error condition can be used to determine the optimal model. $n\rho$ defines a coding rate which should be as large as possible. Furthermore, such a procedure will allow us to measure the number of bits relative to the hypothesis class which we have selected for our pattern recognition problem.

The sender follows the following procedure to define the set of code problems: (i) the problem generator send the data set $\mathbf{X}^{(1)}$ to sender and receiver; (ii) the sender permutes the object indices of the data in such a way that the new optimal solution $c^{\perp}(\mathbf{X})$ is transformed to $\sigma_j \circ c^{\perp}(\mathbf{X})$. In total, there exist $2^{nH(p_\alpha)}$ with $H(p_\alpha) = -\sum_{1 \leq \nu \leq n} np_\nu \log p_\nu$ many transformations σ. This set of transformations is shared with the receiver which establishes a code.

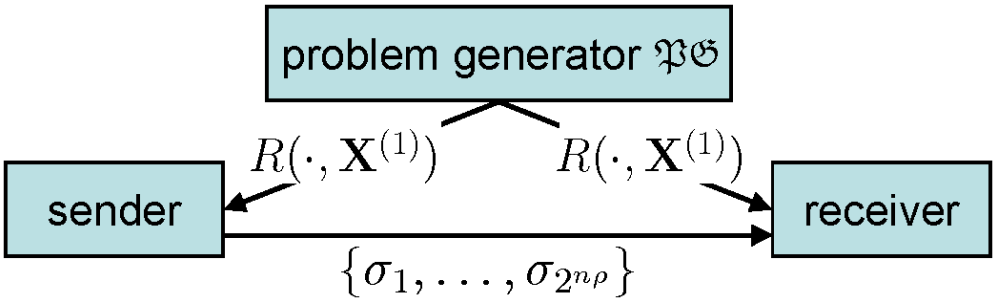

Figure 1: Generation of a set of code problems for communication.

In the **communication process**, the sender selects the transformation σ_s and send this transformation to the problem generator $\mathfrak{PG}$ which generates a second data set $\mathbf{X}^{(2)} \sim \Pr(\mathbf{X})$. This *test* data set is drawn from the same probability distribution as the training data set $\mathbf{X}^{(1)}$. The $\mathfrak{PG}$ then applies the transformation σ_s to the test data and send the transformed data $\tilde{\mathbf{X}} = \sigma_s \circ \mathbf{X}^{(2)}$ to the receiver.

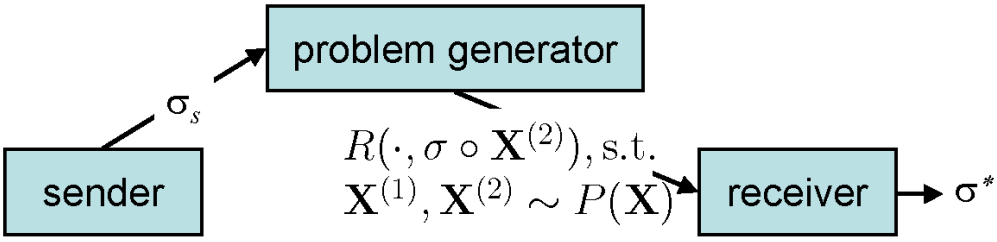

Figure 2: Communication process: the sender selects transformation σ_s and the receiver estimates $\sigma^\star$.

The receiver now faces the question which transformation $\sigma_j,\ 1 \leq j \leq 2^{n\rho}$ has been selected by the sender. If he is able to estimate the transformation selected by the sender, then he has received $n\rho$ bits in this communication. To compute the intersection between the approximation set of the test problem and the approximation set of one of the code problems, solutions of the test problem have to be mapped to the hypothesis class $\mathcal{C}(\mathbf{X}^{(1)})$. This mapping is denoted by $\phi\ :\ \mathcal{C}(\mathbf{X}^{(2)}) \to \mathcal{C}(\mathbf{X}^{(1)})$.

For **decoding**, the receiver intersects the approximation set $\phi \circ \mathcal{C}_\gamma(\tilde{\mathbf{X}})$ with all approximation sets in the codebook $\{\mathcal{C}_\gamma(\sigma_j \circ \mathbf{X}^{(1)}), 1 \leq j \leq 2^{n\rho}\}$. Under the condition that a sufficiently large overlap exists, the receiver declares the transformation $\sigma^\star$

$$\begin{aligned} \sigma^\star &= \arg\max_{\sigma} \left| \mathcal{C}_\gamma(\sigma \circ \mathbf{X}^{(1)}) \cap \phi\left(\mathcal{C}_\gamma(\tilde{\mathbf{X}}^{(2)})\right)\right| \\ &\text{if} \quad \frac{|\mathcal{C}_\gamma(\sigma^* \circ \mathbf{X}^{(1)}) \cap \phi\left(\mathcal{C}_\gamma(\tilde{\mathbf{X}}^{(2)})\right)|}{|\mathcal{C}_\gamma(\sigma^* \circ \mathbf{X}^{(1)})|} \geq 1-\epsilon \end{aligned} \tag{2}$$

as being selected by the sender. $\sigma^\star$ is the received message which has been transmitted by an approximate optimization protocol using the problem generator as a channel.

2.4 Error Analysis of this Code

To analyze the error of this communication protocol, we introduce the following events $\mathcal{E}_j$:

$$\mathcal{E}_j = \left\{ \frac{|\mathcal{C}_\gamma(\sigma_j \circ \mathbf{X}^{(1)}) \cap \left(\phi \circ \mathcal{C}_\gamma(\sigma_s \circ \mathbf{X}^{(2)})\right)|}{|\mathcal{C}_\gamma(\sigma_j \circ \mathbf{X}^{(1)})|} \geq 1-\epsilon \right\} \quad 1 \leq j \leq 2^{n\rho} \tag{3}$$

The event $\mathcal{E}_s,\ 0 < \epsilon \ll 1$ corresponds to correct communication with $\sigma^\star = \sigma_s$.
Two types of errors can occur using this communication protocol:

1. The approximation set $\phi \circ \mathcal{C}_\gamma(\tilde{\mathbf{X}}^{(2)})$ does not substantially intersect with the "correct", sender selected approximation set $\mathcal{C}_\gamma(\sigma_s \circ \mathbf{X}^{(1)})$, i.e.,

$$\overline{\mathcal{E}_s} = \left\{ \frac{|\mathcal{C}(\mathbf{X}^{(1)}) \setminus \mathcal{C}_\gamma(\sigma_s \circ \mathbf{X}^{(1)}) \cap \left(\phi \circ \mathcal{C}_\gamma(\sigma_s \circ \mathbf{X}^{(2)})\right)|}{|\mathcal{C}_\gamma(\sigma_s \circ \mathbf{X}^{(1)})|} \geq 1-\epsilon \right\} \quad 1 \leq j \leq 2^{n\rho} \tag{4}$$

2. The approximation set $\phi \circ \mathcal{C}_\gamma(\tilde{\mathbf{X}}^{(2)})$ substantially intersects with an "incorrect" approximation set $\mathcal{C}_\gamma(\sigma_j \circ \mathbf{X}^{(1)}),\ j \neq s$, i.e., the event $\mathcal{E}_j,\ j \neq s$ occurs.

The conditional error of communication if the sender has selected the transformation σ_s:

$$\begin{aligned} P(\text{error}|\sigma_s) &= P(\mathcal{E}_1 \cup \cdots \cup \overline{\mathcal{E}_s} \cup \cdots \cup \mathcal{E}_{2^{n\rho}}|\sigma_s) \\ &\leq P(\overline{\mathcal{E}_s}|\sigma_s) + \sum_{j \neq s} P(\mathcal{E}_j|\sigma_s) \\ &\leq \epsilon + \sum_{j \neq s} \sum_{\mathcal{C}(X^{(1)})} \frac{I_{\{c \in \phi \circ \mathcal{C}_\gamma(\sigma_s \circ X^{(2)})\}}}{|\phi \circ \mathcal{C}_\gamma(\sigma_s \circ X^{(2)})|} I_{\{c \in \mathcal{C}_\gamma(\sigma_j \circ X^{(1)})\}} P(\sigma_j) \\ &\leq \epsilon + (2^{n\rho}-1)\exp\left(-H(\sigma_j)\right) \times \frac{|\phi \circ \mathcal{C}_\gamma(\sigma_s \circ X^{(2)}) \cap \mathcal{C}_\gamma(\sigma_j \circ X^{(1)})|}{|\phi \circ \mathcal{C}_\gamma(\sigma_s \circ X^{(2)})|} \end{aligned} \tag{5}$$

$H(\sigma_j)$ denotes the entropy of the transformation σ_j which is independent of the index j. The exponentially large terms have to vanish for error free communication. This condition which yields the inequality

$$\begin{aligned} n\rho &\leq \log|\phi \circ \mathcal{C}_\gamma(\sigma_s \circ X^{(2)}) \cap \mathcal{C}_\gamma(\sigma_j \circ X^{(1)})| - \log|\phi \circ \mathcal{C}_\gamma(\sigma_s \circ X^{(2)})| - H(\sigma_j) \\ &\equiv \mathcal{I}(\sigma_s, \mathcal{C}_\gamma(\sigma_s \circ X^{(2)})) \end{aligned} \tag{6}$$

Since $H(\sigma_j) = H(\sigma_s)$ we can abbreviate the right hand side of inequality (6) as a mutual information $\mathcal{I}(.,.)$. This inequality defines an approximation capacity for an optimization problem and it allows us to study the number of bits which can be reliably extracted from an optimization problem $R(c, \theta; \mathbf{X})$ in the presence of noise.

Within the SIMBAD project we will study this concept of structure dependent information to measure the influence of invariances when selecting a cost function for pattern recognition. We also will investigate the role of a metric in the space of hypotheses. Beyond metric approaches, the approximation sets can also be defined by algorithms which do not necessarily follow a gradient dynamics.

3 Bayesian clustering models for dissimilarity matrices

The Bayesian clustering approach presented in this work aims 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.

Despite the lack of a vectorial representation, one way to formulate clustering problems of this kind is to define a generative process that produces points in some Euclidean space according to some form of mixture distribution, and then study the distribution of the pairwise distance matrices associated with sets of points. By construction, however, it is clear that any such approach can only generate metric distances, which might lead to the conclusion that this class of models are not very helpful for explaining general dissimilarity matrices.

This gap between explainable metric distances and non-metric matrices outside the accessible modeling space, however, might be bridged by exploiting some inherent invariance principles of clustering models. A first idea of this kind has been presented in [7] for the k-means clustering model which can be reformulated as a matrix partitioning method for squared Euclidean distances between pairs of points. Once we have arrived at this matrix partitioning problem, however, we can forego the requirement on the distances as being of squared Euclidean form, and simply use this method for partitioning any dissimilarity matrix. This strategy naturally brings up questions about the price to be payed for dropping the Euclidean assumption.

The key insight in [7] is that basically no such price has to be paid at all: it could be shown that the partition structure defined by maximizing the associated cost function is both invariant under symmetrizing transformations and under constant additive shifts applied to the dissimilarities. The first property ensures that any input matrix violating the symmetry axiom of a metric space can be safely symmetrized, and the shift invariance property implies that any (potentially symmetrized) matrix can be forced to fulfill the triangle inequality without changing the solution structure. To be exact, the assertion is even stronger in that under this partitioning model, any dissimilarity matrix can be safely transformed such that the dissimilarities fulfill the parallelogram inequality. It follows that the corresponding norm stems from a dot product which, in turn, implies that there exists a set

of points in $\mathbb{R}^n$ which is geometrically arranged in such a way that the transformed dissimilarities are squared Euclidean distances.

The conclusion is that the the k-means cost function (or its dissimilarity-base counterpart) is essentially "blind" against metric violations. Informally, this property means that the clustering model explains both metric and artificially "metricized" matrices equally well, leading to the some parameter estimates in a maximum likelihood framework.

Despite its elegance, the approach in [7] is particularly tailored to a certain cost function and does not provide general insights into the sources and origins of the identified invariances. In this work we will go one step further and reformulate the matrix partitioning problem in a fully probabilistic framework. We show that an important class of mixture models can be viewed as low-rank matrix approximations, 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 [6], which is then extended to a partitioning process on dissimilarity 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.

3.1 Dirichlet cluster process for Gaussian mixtures.

Let $[n] := \{1, \ldots, n\}$. 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 $\mathbb{C}$. 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 $\mathbb{B}_n$ of partitions of the set $[n]$ 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.

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. [6]: fix the number of mixture modes k, generate mixing proportions $\pi = (\pi_1, \ldots, \pi_k)$ from an exchangeable Dirichlet distribution 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 P of $[n]$ induced by x. Integrating out π, one arrives at a Dirichlet-Multinomial-type prior over partitions $P(B|\lambda, k)$. The limit as $k \to \infty$ is well defined and known as the Ewens process (a.k.a. Chinese Restaurant process), [3]. 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

$$\Sigma_B = I_n \otimes \Sigma_0 + B \otimes \Sigma_1, \quad \text{with} \quad \text{cov}(\mathcal{Y}_{ir}, \mathcal{Y}_{js}|B) = \delta_{ij}\Sigma_{0rs} + B_{ij}\Sigma_{1rs}, \tag{7}$$

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 $\Sigma_B = \alpha I_n + \beta B$. 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 observed data and partitions reads

$$p(Y, B|\alpha, \beta, \lambda, k) = \left[\prod_{b\in B}\prod_{j=1}^{d} N(Y_{i_b j}|\alpha I_{n_b} + \beta E_{n_b})\right] \cdot P(B|\lambda, k), \tag{8}$$

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

3.2 Inner products and distances.

The above concept of a Gauss-Dirichlet cluster process can be easily extended to a sequence of inner product (and distance) matrices. Conditioned on the partition B, the inner product matrix $\mathcal{S} = \mathcal{Y}\mathcal{Y}^t$ follows a (possibly degenerate) Wishart distribution in d degrees of freedom, $\mathcal{S} \sim \mathcal{W}_d(\Sigma_B)$. Regarding the distribution of the distance matrix with components $D_{ij} = S_{ii} + S_{jj} - 2S_{ij}$, it is convenient to first consider the case $d = 1$ and to introduce the notion of a generalized Gaussian distribution with kernel $\mathbb{K} \subset \mathbb{R}^n$, see [5]: $\mathcal{Y} \sim N(\mathbb{K}, \boldsymbol{\mu}, \Sigma)$. Consider a matrix K whose columns are the vectors in $\mathbb{K}$ so that $\mathbb{K} =$ span(K). For any transformation L with $LK = 0$, the meaning of the general Gaussian notation is:

$$L\mathcal{Y} \sim N(L\boldsymbol{\mu}, L\Sigma L^t). \tag{9}$$

In the following we will only consider the case $\boldsymbol{\mu} = \mathbf{0}$ and only such transformation L under which $L\Sigma L^t$ is strict positive definite. It follows that under the kernel $\mathbb{K}$, two parameter settings Σ_1 and Σ_2 are equivalent if $(\Sigma_1 - \Sigma_2) \in$ sym$^2(\mathbb{K} \otimes \mathbb{R}^n)$. It is also useful to introduce the distributional symbol $S \sim \mathcal{W}_1(\mathbb{K}, \Sigma)$ for the generalized Wishart distribution of the random matrix $\mathcal{S} = \mathcal{Y}\mathcal{Y}^t$ when $\mathcal{Y} \sim N(\mathbb{K}, \mathbf{0}, \Sigma)$.

The key observation in [5] is that $D_{ij} = S_{ii} + S_{jj} - 2S_{ij}$ defines a linear transformation on symmetric matrices with kernel sym$^2(\mathbf{1} \otimes \mathbb{R}^n)$ which implies that for $\mathbb{K} = \mathbf{1}$ (the space of constant functions), the distances follow a generalized Wishart distribution: $-\mathcal{D} \sim \mathcal{W}_1(\mathbf{1}, 2\Sigma_B) = \mathcal{W}_1(\mathbf{1}, \Delta)$ with $\Delta_{ij} = \Sigma_{Bii} + \Sigma_{Bjj} - 2\Sigma_{Bij}$. For the above setting with spherical covariances, $\Sigma_0 = \alpha I, \Sigma_1 = \beta I$, this result generalizes in a natural way to d dimensions: $-\mathcal{D} \sim \mathcal{W}_d(\mathbf{1}, \Delta)$.

The restriction $\boldsymbol{\mu} = \mathbf{0}$ does not limit the generality of the approach, since by "centering" the distance matrix we can always transform the representation into inner product form via $S = -\frac{1}{2}QDQ$ with $Q_{ij} = (1 - \frac{1}{n})\delta_{ij}$, which eliminates any contribution of the mean vector while preserving the distances. Note that Q itself is a projection with the same kernel $\mathbf{1}$. Thus, given an observed matrix D, we can directly work with the transformed $S = -\frac{1}{2}QDQ$ assuming the model $S \sim \mathcal{W}_d(\Sigma_B)$. The joint distribution of inner-product matrices and partitions reads

$$p(S, B|\alpha, \beta, \lambda, k) = \mathcal{W}_d(S|(\alpha I_n + \beta B)) \cdot P(B|\lambda, k), \tag{10}$$

which again is invariant under index permutations and factorizes over blocks $b \in B$:

$$p(S, B|\alpha, \beta, \lambda, k) = f(S)\left[\prod_{b\in B} |\Sigma_b|^{-\frac{d}{2}} \exp\left(-\mathrm{tr}(\Sigma_b^{-1}S_b)\right)\right] \cdot P(B|\lambda, k), \tag{11}$$

where $f(S)$ captures terms that are independent of the partition, and Σ_b, S_b denote the submatrices corresponding to the b-th block.

3.3 Relation to multi-dimensional scaling.

Classical multi-dimensional scaling [2] can be interpreted as using a distance model

$$-D \sim \mathcal{W}(\mathbf{1}, \Delta) \text{ with } \Delta = \Delta_0 + M + \sigma^2 I, \tag{12}$$

where Δ_0 stems from the kernel, 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 [5]. It has been proposed to use the transformation $S = -\frac{1}{2}QDQ$ with $Q = (1 - \frac{1}{n})I_n$ (see above), which eliminates Δ_0 and transforms the data into inner product form: $S \sim \mathcal{W}(\mathbf{1}, \Sigma)$, with $\Sigma = \frac{1}{2}\sigma^2 Q + M'$. Note that Q is a scaled identity matrix which means that this expression is essentially the same as our covariance model $\Sigma_B = \alpha I_n + \beta B$. The only difference is that B is not an arbitrary low-rank matrix, but additionally constrained to be a binary partition matrix. Thus, Gaussian mixture models can be understood as a binarized version of multi-dimensional scaling in that the inner product matrix is approximated by a low-rank partition matrix. The white noise term αI_n corresponding to the within-class covariance has the role of absorbing the deviations from the low-rank model.

The expected value of $\mathcal{S} \sim \mathcal{W}_d(\Sigma_B)$ is $E[\mathcal{S}] = \Sigma_B$. Adding an additional noise term δI to the covariance matrix shifts the expected value of $\mathcal{S}$ 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 between a set of n vectors. We will study shift invariance more formally in the next section. In particular, we will show that in the large sample size limit (or alternatively in the hard clustering limit) the mixture model becomes shift invariant.

3.4 Shift invariance.

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

$$\frac{1}{\alpha}\left(I - \frac{\beta}{\alpha + n_b\beta} E_{n_b}\right). \tag{13}$$

Denoting by A the argument of the exponential function in (11), a shift $S' = S + \delta I_n$ implies

$$\alpha A' := -\alpha \operatorname{tr}((\alpha I_{n_b} + \beta E_{n_b})^{-1}(S_{n_b} + \delta I)) = \frac{n_b\beta}{\alpha + n_b\beta}(n_b\bar{S}_{n_b} + \delta) - \operatorname{tr}(S_{n_b}) - n_b\delta, \tag{14}$$

where $\bar{S}_{n_b}$ denotes the mean value of the b-th block of S. For α, β fixed and large block sizes n_b it follows that

$$\begin{aligned} \alpha A' &\approx n_b\bar{S}_{n_b} - \operatorname{tr}(S_{n_b}) - (n_b - 1)\delta \\ \Rightarrow \sum_{b\in B} \alpha A' &\approx -\operatorname{tr}(S) - (n - \#B)\delta + \sum_{b\in B} n_b\bar{S}_{n_b}, \end{aligned} \tag{15}$$

with $\#B$ being the number of blocks in the partition B. This result implies that for fixed parameters and number of blocks, the conditional posterior of partitions is approximately shift invariant. In the limit $n \to \infty$ or in the hard clustering limit $\beta \to \infty$ this statement becomes exact.

3.5 Conclusion.

Gaussian mixture models can be defined via an exchangeable partitioning process which avoids problems related to unidentifiable of labels. Given a partition, any set of n observations is generated according to a Gaussian distribution. Assuming spherical within- and between covariances, this construction induces a Wishart model for inner product matrices which is formally identical to a probabilistic version of (classical) multi-dimensional scaling. This analogy shows that mixture models can be viewed as low-rank matrix approximations using binary partition matrices. The within-class covariance term has the role of a white noise term capturing deviations from the low-rank model.

A partitioning model on dissimilarity matrices 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 preference for partitions. For a fixed number of observations n, certain models of this kind can be formally described as mixture models in $\mathbb{R}^n$, even if the input matrix violates the triangle inequality.

We have shown that the clustering model induced by an exchangeable partitioning process and a Wishart model on inner product matrices is "nearly" shift invariance, a statement which becomes exact if either the block sizes go to infinity, or in the hard-clustering limit. On an intuitive level, shift invariance can be explained as "absorbing" shifts in the white-noise term that captures deviations from the low-rank partition model.

4 Summary

The work of the ETH Zurich group together with the group of Volker Roth (Univ. Basel) has focussed on both the theoretical studies of clustering and on the data analysis part of the medical application in cancer studies. The theoretical work is concerned with model selection in clustering and it pursues two strategies, (i) an information theoretic approach to measure the information content of hypothesis classes in an approximation setting and (ii) a non-parametric Bayesian approach to understand pairwise clustering and its shift invariance. Both studies are complementary and will be brought together in the next stage of the project.

References

[1] Joachim M. Buhmann. Empirical risk approximation: An induction principle for unsupervised learning. Technical Report IAI-TR-98-3, Department of Computer Science III / University of Bonn, 1998.

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

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

[4] Tilman Lange, Mikio Braun, Volker Roth, and Joachim M. Buhmann. Stability-based validation of clustering solutions. *Neural Computation*, 16(6):1299–1323, June 2004.

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

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

[7] 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.