The Translation-invariant Wishart-Dirichlet Process for Clustering Distance Data

Julia E. Vogt JULIA.VOGT@UNIBAS.CH
Computer Science Department, University of Basel, Basel, Switzerland

Sandhya Prabhakaran SANDHYA.PRABHAKARAN@UNIBAS.CH
Computer Science Department, University of Basel, Basel, Switzerland

Thomas J. Fuchs THOMAS.FUCHS@INF.ETHZ.CH
Department of Computer Science, ETH Zurich & Competence Center for Systems Physiology and Metabolic Diseases, Zurich, Switzerland

Volker Roth VOLKER.ROTH@UNIBAS.CH
Computer Science Department, University of Basel, Basel, Switzerland

Abstract

We present a probabilistic model for clustering of objects represented via pairwise dissimilarities. We propose that even if an underlying vectorial representation exists, it is better to work directly with the dissimilarity matrix hence avoiding unnecessary bias and variance caused by embeddings. By using a Dirichlet process prior we are not obliged to fix the number of clusters in advance. Furthermore, our clustering model is permutation-, scale- and translation-invariant, and it is called the Translation-invariant Wishart Dirichlet (TIWD) process. A highly efficient MCMC sampling algorithm is presented. Experiments show that the TIWD process exhibits several advantages over competing approaches.

1. Introduction

The Bayesian clustering approach presented in this work aims at identifying subsets (or "clusters") of objects represented as columns/rows in a 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. Such underlying vectorial representation may or may not exist, depending on whether the dissimilarity matrix can be embedded (without distortion) in a vector space. One way of dealing with such problems would be to explicitly construct an Euclidean embedding (or possibly a distorted embedding), and to apply a traditional clustering method in the Euclidean space. We argue, however, that even under the assumption that there exists an Euclidean embedding it is better *not* to embed the data, since any such choice might induce an unnecessary bias and variance in the clustering process. Technically speaking, such embeddings break the symmetry induced by the translation- and rotation-invariance which reflects the information loss incurred when moving from vectors to pairwise dissimilarities. We propose a clustering model which works directly on dissimilarity matrices. It is invariant against label- and object permutations and against scale transformations. The model is fully probabilistic in nature, which means that on output we are given samples from a distribution over partitions. Further, the use of a Dirichlet process prior unburdens the user from explicitly fixing the number of clusters. We present a highly efficient sampling algorithm which avoids costly matrix operations by carefully exploiting the structure of the clustering problem. Invariance against label permutations is a common cause of the so-called "label switching" problem in mixture models, (Jasr et al., 2005). By formulating the model as a partition process this switching problem is circumvented.

This paper is structured as follows: we start with a review of the *Dirichlet cluster process for Gaussian mixtures* in (McCullagh & Yang, 2008). This model is generalized to relational data by enforcing translation invariance. We call this new model the *Translation-invariant Wishart-*

Appearing in *Proceedings of the 27^{th} International Conference on Machine Learning*, Haifa, Israel, 2010. Copyright 2010 by the author(s)/owner(s).

Dirichlet (WD) cluster process. We then develop an efficient sampling algorithm which makes it possible to apply the method to large-scale datasets.

2. Gauss-Dirichlet Cluster Process

Let $[n] := \{1, \dots, n\}$ 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 $y(i) = y(j)$ and $B(i, j) = 0$ otherwise, with y 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 (Pitman, 2006).

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 (MacEachern, 1994; Dahl, 2005; McCullagh & Yang, 2008): fix the number of mixture modes k, generate mixing proportions $\pi = (\pi_1, \dots, \pi_k)$ from an exchangeable Dirichlet distribution $\text{Dir}(\xi/k, \dots, \xi/k)$, generate a label sequence $\{y(1), \dots, y(n)\}$ from a multinomial distribution and forget the labels introducing the random partition B of $[n]$ induced by y. Integrating out π, one arrives at a Dirichlet-Multinomial prior over partitions

$$P_n(B|\xi, k) = \frac{k!}{(k - k_B)!} \frac{\Gamma(\xi) \prod_{b \in B} \Gamma(n_b + \xi/k)}{\Gamma(n + \xi)[\Gamma(\xi/k)]^{k_B}}, \quad (1)$$

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 (Ewens, 1972; Neal, 2000; Blei & Jordan, 2006). Given such a partition B, a sequence of n-dimensional observations $\boldsymbol{x}_i \in \mathbb{R}^n,\ i = 1, \dots, d$ is arranged as columns of the $(n \times d)$ matrix X, and this X is generated from a zero-mean Gaussian distribution with covariance matrix

$$\begin{aligned} \widetilde{\Sigma}_B &= I_n \otimes \Sigma_0 + B \otimes \Sigma_1, \\ \text{with} \quad \text{cov}(X_{ir}, X_{js}|B) &= \delta_{ij}\Sigma_{0rs} + B_{ij}\Sigma_{1rs}, \end{aligned} \quad (2)$$

where Σ_0 is the usual $(d \times d)$ "pooled" within-class covariance matrix and Σ_1 the $(d \times d)$ between-class matrix, respectively, and δ_{ij} denotes the Kronecker symbol. Since the partition process is invariant under permutations, we can always think of B being block-diagonal. For spherical covariance matrices (i.e. scaled identity matrices), $\Sigma_0 = \alpha I_d, \Sigma_1 = \beta I_d$, the covariance structure reduces to

$$\begin{aligned} \widetilde{\Sigma}_B &= I_n \otimes \alpha I_d + B \otimes \beta I_d \\ &= (\alpha I_n + \beta B) \otimes I_d =: \Sigma_B \otimes I_d, \\ \text{with} \quad \text{cov}(X_{ir}, X_{js}|B) &= (\alpha\delta_{ij} + \beta B_{ij})\delta_{rs}. \end{aligned} \quad (3)$$

Thus, the columns of X are independent n-dimensional vectors $\boldsymbol{x}_i \in \mathbb{R}^n$ distributed according to a normal distribution with covariance matrix $\Sigma_B = \alpha I_n + \beta B$. Further, the distribution factorizes over the blocks $b \in B$. Introducing the symbol $i_b := \{i : i \in b\}$ defining an index-vector of all objects assigned to block b, the joint distribution reads

$$\begin{aligned} p(X, B|\alpha, \beta, \xi, k) &= P_n(B|\xi, k) \\ &\cdot \left[\prod_{b \in B} \prod_{j=1}^{d} N(X_{i_b j}|\alpha I_{n_b} + \beta \mathbf{1}_{n_b}\mathbf{1}_{n_b}^t)\right], \end{aligned} \quad (4)$$

where n_b is the size of block b and $\mathbf{1}_{n_b}$ a n_b-vector of ones. In the following we will use the abbreviations $\mathbf{1}_b := \mathbf{1}_{n_b}$ and $I_b := I_{n_b}$ to avoid double subscripts. Note that this distribution is expressed in terms of the partition without resorting to labels, so label switching cannot occur.

3. 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 $X_{n \times d}$ follows the zero-mean Gaussian distribution specified in (2), with $\Sigma_0 = \alpha I_d, \Sigma_1 = \beta I_d$. Then, conditioned on the partition B, the inner product matrix $S = XX^t/d$ follows a (possibly singular) Wishart distribution in d degrees of freedom, $S \sim \mathcal{W}_d(\Sigma_B)$, (Srivastava, 2003). If we directly observe the dot products S, 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:

$$\begin{aligned} P_n(B|S, \alpha, \beta, \xi, k) &\propto \mathcal{W}_d(S|\Sigma_B) \cdot P_n(B|\xi, k) \\ &\propto |\Sigma_B|^{-\frac{d}{2}} \exp\left(-\tfrac{d}{2}\text{tr}(\Sigma_B^{-1}S)\right) \cdot P_n(B|\xi, k), \end{aligned} \quad (5)$$

For the following derivation it is suitable to re-parametrize the model in terms of (α, θ) instead of (α, β), where $\theta := \beta/\alpha$, and in terms of $W := \Sigma_B^{-1}$. Due to the block structure in B, $P_n(B|S, \bullet)$ factorizes over the blocks $b \in B$:

$$\begin{aligned} P_n(B|S, \alpha, \theta, \xi, k) &\propto P_n(B|\xi, k) \\ &\cdot \left[\prod_{b \in B} |W_b|^{\frac{d}{2}}\right] \exp\left(-\sum_{b \in B} \tfrac{d}{2}\text{tr}(W_b S_{bb})\right), \end{aligned} \quad (6)$$

where W_b, S_{bb} denote the submatrices corresponding to the b-th diagonal block in B or W, see Figure 1.

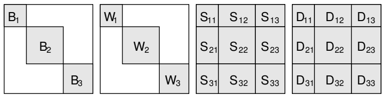

Figure 1. Example of the block structure of B and W (left) and the definition of sub-matrices in S and D (right) for $k_B = 3$.

The above factorization property can be exploited to derive an efficient inference algorithm for this model. The key observation is that the inverse matrix $W_b = \Sigma_b^{-1}$ can be analytically computed as

$$W_b = (\alpha I_b + \beta \mathbf{1}_b \mathbf{1}_b^t)^{-1} = \frac{1}{\alpha}\left[I_b - \frac{\theta}{1 + n_b\theta}\mathbf{1}_b\mathbf{1}_b^t\right]. \quad (7)$$

Thus, the contribution of block b to the trace is

$$\mathrm{tr}(W_b S_{bb}) = \frac{1}{\alpha}\left[\mathrm{tr}(S_{bb}) - \frac{\theta}{1+n_b\theta}\bar{S}_{bb}\right], \tag{8}$$

where $\bar{S}_{bb} = \mathbf{1}_b^t S_{bb}\mathbf{1}_b$ denotes the sum of the b-th diagonal block of S. A similar trick can be used for the determinant which is the product of the eigenvalues: the k_B smallest eigenvalues of W are given by $\lambda_b = \alpha^{-1}(1+\theta n_b)^{-1}$. The remaining $n - k_B$ eigenvalues are equal to α^{-1}. Thus, the determinant reads

$$|W| = \prod_{b\in B}\lambda_b = \alpha^{-n}\prod_{b\in B}(1+\theta n_b)^{-1}. \tag{9}$$

3.1. Scale Invariance

Note that the re-parametrization using (α, θ) leads to a new semantics of $(1/\alpha)$ as a scale parameter: we excluded α from the partition-dependent terms in the product over the blocks in (9), which implies that the conditional for the partition becomes

$$\begin{aligned} &P_n(B|\bullet) \propto P_n(B|\xi, k)\cdot\left[\prod_{b\in B}(1+\theta n_b)^{-1}\right]^{-d/2} \\ &\cdot\exp\left(-\frac{1}{\alpha}\frac{d}{2}\sum_{b\in B}\mathrm{tr}(W_b S_{bb})\right). \end{aligned} \tag{10}$$

Note that $(1/\alpha)$ simply rescales the observed matrix S, and we can make the model scale invariant by introducing a prior distribution and integrating out α. The conditional posterior for α follows an inverse Gamma distribution

$$p(\alpha|r,s) = \frac{s^r}{\Gamma(r)}\left(\frac{1}{\alpha}\right)^{r+1}\exp\left(-\frac{s}{\alpha}\right), \tag{11}$$

with shape parameter $r = n\cdot d/2 - 1$ and scale $s = \frac{d}{2}(\mathrm{tr}(S) - \sum_{b\in B}\frac{\theta}{1+n_b\theta}\bar{S}_{bb})$, cf. eqs. (8) and (10). Using an inverse Gamma prior with parameters r_0, s_0, the posterior is of the same functional form with $r_p = r + r_0 + 1$ and $s_p = s + s_0$, and we can integrate out α analytically. Dropping all terms independent of the partition structure we arrive at

$$P_n(B|\bullet) \propto P_n(B|\xi, k)|W|_{(\alpha=1)}^{d/2}(s+s_0)^{r+r_0+1}, \tag{12}$$

where $|W|_{(\alpha=1)} = \prod_{b\in B}(1+\theta n_b)^{-1}$ follows from (9).

3.2. The Centering Problem

In practice, however, there are two problems with the model described above: (i) we often do not directly observe S, but only a matrix of *distances* D. In the following we will assume that the (suitably pre-processed) matrix D contains *squared Euclidean distances* with components $D_{ij} = S_{ii} + S_{jj} - 2S_{ij}$; (ii) even if we observe a dot-product matrix, we usually have no information about the mean vector $\boldsymbol{\mu}$. Note that we assumed that there exists a matrix X with $XX^t = S$ such that the *columns* of X are independent copies drawn from a zero-mean Gaussian in $\mathbb{R}^n$: $\boldsymbol{x} \sim N(\boldsymbol{\mu} = \mathbf{0}_n, \Sigma = \Sigma_B)$. This assumption is crucial, since general mean vectors correspond to a *noncentral* Wishart model (Anderson, 1946), which imposes severe computational problems due to the appearance of the hypergeometric function. Both of the above problems are related in that they have to do with the lack of information about geometric transformations: assume we only observe S without access to the vectorial representations $X_{n\times d}$. Then we have lost the information about orthogonal transformations $X \leftarrow XO$ with $OO^t = I_d$, i.e. about rotations and reflections of the rows in X. If we only observe D, we have additionally lost the information about translations of the rows. Our sampling model implies that the means in each row are expected to converge to zero as the number of replications d goes to infinity. Thus, if we had access to X and if we are not sure that the above zero-mean assumption holds, it might be a plausible strategy to subtract the empirical row means, $X_{n\times d} \leftarrow X_{n\times d} - (1/d)X_{n\times d}\mathbf{1}_d\mathbf{1}_d^t$, and then to construct a candidate matrix S by computing the pairwise dot products. This procedure should be statistically robust if $d \gg n$, since then the empirical means are probably close to their expected values. Such a matrix S fulfills two requirements for selecting candidate dot product matrices: first, S should be "typical" with respect to the assumed Wishart model with $\boldsymbol{\mu} = \mathbf{0}$, thereby avoiding any bias introduced by a particular choice. Second, the choice should be robust in a statistical sense: if we are given a second observation from the same data source, the two selected prototypical matrices S_1 and S_2 should be similar. For small d, this procedure is dangerous since it can introduce a strong bias even if the model is correct.

Consider now case (ii) where we observe S without access to X. Case (i) needs no special treatment, since it can be reduced to case (ii) by first constructing a positive semi-definite matrix S which fulfills $D_{ij} = S_{ii} + S_{jj} - 2S_{ij}$. For "correcting" the matrix S just as described above we would need a procedure which effectively subtracts the empirical row means from the rows of X. Unfortunately, there exists no such matrix transformation that operates directly on S without explicit construction of X. It is important to note that the "usual" centering transformation $S \leftarrow QSQ$ with $Q_{ij} = \delta_{ij} - \frac{1}{n}$ as used in kernel PCA and related algorithms does not work here: in kernel PCA the rows of X are assumed to be i.i.d. replications in $\mathbb{R}^d$. Consequently, the centered matrix S_c is built by subtracting the *column* means: $X_{n\times d} \leftarrow X_{n\times d} - (1/n)\mathbf{1}_n\mathbf{1}_n^t X_{n\times d}$ and $S_c = XX^t = QSQ$. Here, we need to subtract the *row* means, and therefore it is inevitable to explicitly construct X, which implies that we have to choose a certain orthogonal transformation O. It might be reasonable to consider only rotations and to use the principle components as coordinate axes. This is essentially the kernel PCA embedding procedure: compute $S_c = QSQ$ and its eigenvalue decomposition $S_c = V\Lambda V^t$, and then project on the principle axes: $X = V\Lambda^{1/2}$. The problem with this vector-space embedding is that it is statistically robust in the above sense only if d is small, because otherwise the directions of the principle axes might be difficult to estimate, and the estimates for two replicated observations might highly fluctuate, leading to different row-mean normalizations. Note that this condition for fixing the rotation contradicts the above condition $d \gg n$ that justifies the subtraction of the means. Further, row-mean normalization will change the pairwise dissimilarities D_{ij} (even if the model is correct!), and this change can be drastic if d is small.

The cleanest solution might be to consider the dissimilarities D (which are observed in case (i) and computed as $D_{ij} = S_{ii} + S_{jj} - 2S_{ij}$ in case (ii)) as the "reference" quantity, and to avoid an explicit choice of S and X altogether. Therefore, we propose to encode the translation invariance directly into the likelihood, which means that the latter becomes constant on all matrices S that fulfill $D_{ij} = S_{ii} + S_{jj} - 2S_{ij}$.

3.3. The Translation-invariant WD-Process

A squared Euclidean distance matrix D is characterized by the property of being of *negative type*, which means that $\boldsymbol{x}^t D\boldsymbol{x} = -2\boldsymbol{x}^t S\boldsymbol{x} \leq 0$ for any $\boldsymbol{x} : \boldsymbol{x}^t\mathbf{1} = 0$. This condition is equivalent to the absence of negative eigenvalues in $S_c = QSQ = -(1/2)QDQ$. The distribution of D has been formally studied in (McCullagh, 2009), where it was shown that if S follows a standard Wishart generated from an underlying zero-mean Gaussian process, $S \sim \mathcal{W}_d(\Sigma_B)$, $-D$ follows a generalized Wishart distribution, $-D \sim \mathcal{W}(\mathbf{1}, 2\Sigma_B) = \mathcal{W}(\mathbf{1}, -\Delta)$ defined with respect to the transformation kernel $\mathbb{K} = \mathbf{1}$, where $\Delta_{ij} = \Sigma_{B\,ii} + \Sigma_{B\,jj} - 2\Sigma_{B\,ij}$. To understand the role of the transformation kernel it is useful to introduce the notion of a generalized Gaussian distribution with kernel $\mathbb{K} = \mathbf{1}$: $X \sim N(\mathbf{1}, \boldsymbol{\mu}, \Sigma)$. For any transformation L with $L\mathbf{1} = \mathbf{0}$, the meaning of the general Gaussian notation is:

$$LX \sim N(L\boldsymbol{\mu}, L\Sigma L^t). \tag{13}$$

It follows that under the kernel $\mathbb{K} = \mathbf{1}$, two parameter settings $(\boldsymbol{\mu}_1, \Sigma_1)$ and $(\boldsymbol{\mu}_2, \Sigma_2)$ are equivalent if $L(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \mathbf{0}$ and $L(\Sigma_1 - \Sigma_2)L^t = 0$, i.e. if $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \in \mathbf{1}$, and $(\Sigma_1 - \Sigma_2) \in \{\mathbf{1}_n\boldsymbol{v}^t + \boldsymbol{v}\mathbf{1}_n^t : \boldsymbol{v} \in \mathbb{R}^n\}$, a space which is usually denoted by $\mathrm{sym}^2(\mathbf{1} \otimes \mathbb{R}^n)$. It is also useful to introduce the distributional symbol $S \sim \mathcal{W}(\mathbb{K}, \Sigma)$ for the generalized Wishart distribution of the random matrix $S = XX^t$ when $X \sim N(\mathbb{K}, \mathbf{0}, \Sigma)$. The key observation in (McCullagh, 2009) is that $D_{ij} = S_{ii} + S_{jj} - 2S_{ij}$ defines a linear transformation on symmetric matrices with kernel $\mathrm{sym}^2(\mathbf{1} \otimes \mathbb{R}^n)$ which implies that the distances follow a generalized Wishart distribution with kernel $\mathbf{1}$: $-D \sim \mathcal{W}(\mathbf{1}, 2\Sigma_B) = \mathcal{W}(\mathbf{1}, -\Delta)$. In the multi-dimensional case with spherical within- and between covariances we generalize the above model to Gaussian random matrices $X \sim N(\boldsymbol{\mu}, \Sigma_B \otimes I_d)$. Note that the d columns of this matrix are i.i.d. copies. The distribution of the matrix of squared Euclidean distances D then follows a generalized Wishart with d degrees of freedom $-D \sim \mathcal{W}_d(\mathbf{1}, -\Delta)$. This distribution differs from a standard Wishart in that the inverse matrix $W = \Sigma_B^{-1}$ is substituted by the matrix $\widetilde{W} = W - (\mathbf{1}^t W\mathbf{1})^{-1} W\mathbf{1}\mathbf{1}^t W$ and the determinant $|\cdot|$ is substituted by a generalized $\det(\cdot)$-symbol which denotes the product of the nonzero eigenvalues of its matrix-valued argument (note that $\widetilde{W}$ is rank-deficient). The conditional probability of a partition then reads

$$\begin{aligned} P(B|D, \bullet) &\propto \mathcal{W}_d(-D|\mathbf{1}, -\Delta) \cdot P_n(B|\xi, k) \\ &\propto \det(\widetilde{W})^{\frac{d}{2}} \exp\left(\tfrac{d}{4}\mathrm{tr}(\widetilde{W}D)\right) \cdot P_n(B|\xi, k). \end{aligned} \tag{14}$$

Note that in spite of the fact that this probability is written as a function of $W = \Sigma_B^{-1}$, it is constant over all choices of Σ_B which lead to the same Δ, i.e. independent under translations of the row vectors in X. For the purpose of inferring the partition B, this invariance property means that we can simply use our block-partition covariance model Σ_B and assume that the (unobserved) matrix S follows a standard Wishart distribution parametrized by Σ_B. We do not need to care about the exact form of S, since the conditional posterior for B depends only on D.

Scale invariance can be built into the model with the same procedure as described above for the simple (i.e. not translation invariant) WD-process. The posterior of α again follows an inverse Gamma distribution, and after introducing a prior with parameters (s_0, r_0) and integrating out α we arrive at an expression analogous to (12) with $s = \frac{d}{4}\mathrm{tr}(\widetilde{W}D)$:

$$P(B|\bullet) \propto P_n(B|\xi, k) \det(\widetilde{W}_{(\alpha=1)})^{\frac{d}{2}} \, (s+s_0)^{n\frac{d}{2}+r_0}. \tag{15}$$

3.4. Efficient Inference via Gibbs Sampling

In Gibbs sampling one iteratively samples parameter values from the full conditionals. Our model includes the following parameters: the partition B, the scale α, the covariance parameter θ, the number k of clusters in the population, the Dirichlet rate ξ and the degrees of freedom d. We propose to fix d, ξ and k: the **degrees of freedom** d might be estimated by the rank of S, which is often known from a pre-processing procedure. Note that d is not a very critical parameter, since all likelihood contributions are basically raised to the power of d. Thus, d might be used as an annealing-type parameter for "freezing" a representative partition in the limit $d \to \infty$. Concerning the **number** k **of clusters in the population**, there are two possibilities. Either one assumes $k = \infty$, which results in the Ewens-process model, or one expects a finite k. Our framework is applicable to both scenarios. Estimation of k, however is nontrivial if no precise knowledge about ξ is available. Unfortunately, this is usually the case, and $k = \infty$ might be a plausible assumption in many applications. Alternatively, one might fix k to a large constant which serves as an upper bound of the expected number, which can be viewed as truncating the Ewens process. The **Dirichlet rate** ξ is difficult to estimate, since it only weakly influences the likelihood. Consistent ML-estimators only exist for $k = \infty$: $\hat{\xi} = k_B/\log n$, and even in this case the variance only decays like $1/\log(n)$, cf. (Ewens, 1972). In practice, we should not expect to be able to reliably estimate ξ. Rather, we should have some intuition about ξ, maybe guided by the observation that under the Ewens process model the probability of two objects belonging to the same cluster is $1/(1+\xi)$. We can then either define an appropriate prior distribution, or we can fix ξ. Due to the weak effect of ξ on conditionals, these approaches are usually very similar.

The scale α can be integrated out analytically. The likelihood in θ is not of recognized form, and we propose to use a discretized prior set $\{p(\theta_j)\}_{j=1}^{J}$ for which we compute the posteriors $\{p(\theta_j|\bullet)\}_{j=1}^{J}$. A new value of θ is then sampled from the categorical distribution defined by $\{p(\theta_j|\bullet)\}_{j=1}^{J}$. We define a *sweep* of the Gibbs sampler as one complete update of (B, θ). The most time consuming part in a sweep is the update of B by re-estimating the assignments to blocks for a single object (characterized by a row/column in D), given the partition of the remaining objects. Therefore we have to compute the membership probabilities in all existing blocks (and in a new block) by evaluating equation (15), which looks formally similar to (12), but a factorization over blocks is no longer obvious. Every time a new partition is analyzed, a naive implementation requires $O(n^3)$ costs for computing the determinant of $\widetilde{W}$ and the product $\widetilde{W}D$. In one sweep we need to compute k_B such probabilities for n objects, summing up to costs of $O(n^4 k_B)$. However, a more efficient algorithm exists:

Theorem 1. *Assuming k_B blocks in the actual partition and a fixed maximum iteration number in numerical root-finding, a sweep of the Gibbs sampler for the translation-invariant WD model can be computed in $O(n^2 + nk_B^2)$ time.*

Proof. Assume we want to compute the membership probabilities of the l-th object, given the partition of the remaining objects and all other parameter values. We first have to downdate all quantities which depend on object l and its current block and compute the assignment probabilities for all blocks (and a new one). From the resulting categorical distribution we sample a new assignment (say block c) and update all quantities depending on object l and block c. We repeat this procedure for all objects $l = 1, \dots, n$. Since up- and down-dating are reverse to each other but otherwise identical operations, it suffices to consider the updating situation. To compute the membership probabilities we have to assign the new object to a block and evaluate (15) for the augmented matrix D_*, which has one additional column and row. For notational simplicity we will drop the subscript $_*$. Eq. (15) has two components: the prior $P(B|\xi, k)$ and the likelihood term which requires us to compute $\det(\widetilde{W}_{(\alpha=1)})$ and tr$(\widetilde{W}D)$. With identity $\Gamma(x+1) = x\Gamma(x)$ in (1), the contribution of the prior is $n_c + \xi/k$ for existing clusters and $\xi(1 - k_B/k)$ for a new cluster (one simply sets $k = \infty$ for the Ewens-process).

For the likelihood term, consider first the generalized determinant $\det(\widetilde{W})$ in (15). Since $\widetilde{W} = W - (\mathbf{1}^t W\mathbf{1})^{-1} W\mathbf{1}\mathbf{1}^t W$, we have to compute $\rho := (\mathbf{1}^t W\mathbf{1})^{-1}$ for the augmented matrix W after assigning the new object l to block c. Analyzing (7) one derives $\rho^{-1} = \sum_{b\in B} n_b\lambda_b$, where $\lambda_b = (1+\theta n_b)^{-1}$ are the k_B smallest eigenvalues of $W_{(\alpha=1)}$, see eq. (9). Thus, we increase n_c, recompute λ_c and update ρ. Given ρ, we need to compute the eigenvalues of $W - \rho W\mathbf{1}\mathbf{1}^t W =: W - \rho\boldsymbol{v}\boldsymbol{v}^t$, where the latter term defines a rank-one update of W. Analyzing the characteristic polynomial, it is easily seen that the (size-ordered) eigenvalues $\tilde{\lambda}_i$ of $\widetilde{W}$ fulfill three conditions, see (Golub & Van Loan, 1989): (i) the smallest eigenvalue is zero: $\tilde{\lambda}_1 = 0$; (ii) the largest $n - k_B$ eigenvalues are identical to their counterparts in W: $\tilde{\lambda}_i = \lambda_i,\ i = k_B+1, \dots, n$; (iii) for the remaining eigenvalues with indices $i = 2, \dots, k_B$ it holds that if λ_i is a repeated eigenvalue of W, $\tilde{\lambda}_i = \lambda_i$. Otherwise, they are the simple roots of the *secular* equation $f(y) = \rho + \sum_{j=1}^{k_B} \frac{n_j\lambda_j^2}{y-\lambda_j}$ fulfilling the relations $\lambda_i < \tilde{\lambda}_{i+1} < \lambda_{i+1}$. Note that f can be evaluated in $O(k_B)$ time, and with a fixed maximum number of iterations in the root-finding procedure, $\det(\widetilde{W})$ can be computed in $O(k_B)$. A sweep involves n "new" objects and k_B blocks. Thus, the costs sum up to $O(nk_B^2)$:

```
for $i = 1$ to $n$ do
    for $c = 1$ to $k_B$ do
        $n_c \leftarrow n_c + 1$, recompute $\lambda_c$ and update $\rho \rightsquigarrow O(1)$
        Find roots of secular equation $\rightsquigarrow O(k_B)$
    end for
end for
```

For the trace tr$(\widetilde{W}D)$ we have to compute

$$\begin{aligned}\text{tr}(\widetilde{W}D) &= \text{tr}(WD) - \rho\cdot\text{tr}(W\mathbf{1}\mathbf{1}^t WD)\\ &= \text{tr}(WD) - \rho\cdot\mathbf{1}^t WDW\mathbf{1}.\end{aligned} \tag{16}$$

We first precompute $\forall a \in B$: $\bar{D}_{ia} = \sum_{j\in a} D_{ij}$, which induces $O(n)$ costs since there are n summations in total. The first term in (16) is tr$(WD) = \sum_{b\in B}$ tr$(D_{bb}) - \frac{\theta}{1+n_b\theta}\bar{D}_{bb}$, so we first update $\bar{D}$ by recomputing its c-th row/column: update $\gamma_c = n_c\lambda_c$ and $\forall a \in B : \bar{D}_{ac} \leftarrow \bar{D}_{ac} + \bar{D}_{ia} + D_{ii}\delta_{a,c} \rightsquigarrow O(k_B)$ time, and update the c-

th term in tr(WD) in constant time. Defining $\bar{D}_{ab} := \mathbf{1}_a^t D_{ab} \mathbf{1}_b$ and $\gamma_a := \frac{n_a\theta}{1+n_a\theta}$, the second term in (16) reads

$$\begin{aligned} &\rho \textstyle\sum_{ab\in B} \mathbf{1}_a^t W_a D_{ab} W_b \mathbf{1}_b =: \rho \textstyle\sum_{ab\in B} \overline{\Phi}_{ab}, \\ &\overline{\Phi}_{ab} = \bar{D}_{ab} - \gamma_a \bar{D}_{ab} - \gamma_b \bar{D}_{ab} + \gamma_a \gamma_b \bar{D}_{ab}. \end{aligned} \quad (17)$$

Since we have already updated γ and $\bar{D}$, it requires $O(k_B)$ time to update the c-th row. In a sweep, the costs for the trace sum up to $O(n^2 + nk_B^2)$:

```
for $i = 1$ to $n$ do
  $\forall a \in B$: $\bar{D}_{ia} = \sum_{j\in a} D_{ij} \rightsquigarrow O(n)$
  for $c = 1$ to $k_B$ do
    Update $\bar{D} \rightsquigarrow O(k_B)$
    Recompute $c$-th term in tr$(WD) \rightsquigarrow O(1)$
    Compute $\forall a \in B:\ \overline{\Phi}_{ac} = \overline{\Phi}_{ca} \rightsquigarrow O(k_B)$
  end for
end for
```

The sweep is completed by resampling θ from a discrete set with J levels which induces costs of $O(k_B^2)$. □

From the above theorem it follows that the worst case complexity in one sweep is $O(n^3)$ in the infinite mixture (i.e. Ewens process-) model, since $k_B \le n$, and $O(n^2)$ for the truncated Dirichlet process with $k_B \le k < \infty$. If the "true" k is finite, but one still uses the infinite model, it is very unlikely to observe the worst-case $O(n^3)$ behaviour in practice: if the sampler is initialized with a one-block partition (i.e. $k_B = 1$), the trace of k_B typically shows an "almost monotone" increase during burn-in, see Figure 3.

3.5. Model Extensions

One possible extension of the TIWD cluster process is to include some **preprocessing step**. From the model assumptions $S \sim \mathcal{W}(\Sigma_B)$ it follows that if Σ_B contains k_B blocks and if the separation between the clusters (i.e. θ) is not too small, there will be only k_B dominating eigenvalues in S. Thus, one might safely apply kernel PCA to the centered matrix $S_c = -(1/2)QDQ$, i.e. compute $S_c = V\Lambda V^t$, consider only the first $\tilde{k}$ "large" eigenvalues in Λ for computing a low-rank approximation $\tilde{S}_c = V\tilde{\Lambda}V^t$, and switch back to dissimilarities via $\tilde{D}_{ij} = (\tilde{S}_c)_{ii} + (\tilde{S}_c)_{jj} - 2(\tilde{S}_c)_{ij}$. Such preprocessing might be particularly helpful in cases where $S_c = -(1/2)QDQ$ contains some negative eigenvalues which are of relatively small magnitude. Then, the low-rank approximation might be positive semi-definite so that $\tilde{D}$ contains squared Euclidean distances. Such situations occur frequently if the dissimilarities stem from pairwise comparison procedures which can be interpreted as approximations to models which are guaranteed to produce Mercer kernels. A popular example are classical string alignments which might be viewed as approximations of probabilistic alignments using pairwise hidden Markov models. We present such an example in section 4. The downside of kernel PCA are the added costs of $O(n^3)$, but randomized approximation methods have been introduced which significantly reduce these costs. In our TIWD software we have implemented a "symmetrized" version of the random projection algorithm for low-rank matrix approximation proposed in (Vempala, 2004) which uses the idea proposed in (Belabbas & Wolfe, 2007).

Another extension of the model concerns **semi-supervised** situations where for a subset of n_m observations class labels, i.e. assignments to k_m groups, are known. We denote this subset by the set of row indices $\mathbb{A} = \{1, \ldots, n_m\}$. Traditional semi-supervised learning methods assume that at least one labeled object per class is observed, i.e. that the number of classes is known. This assumption, however, is questionable in many real world examples. We overcome this limitation by simply fixing the assignment to blocks for objects in $\mathbb{A}$ during the sampling procedure, and re-estimating only the assignments for the unlabeled objects in $\mathbb{B} = \{n_m + 1, \ldots, n\}$. Using an Ewens process model with $k = \infty$ (or a truncated version thereof with $k > k_m$), the model has the freedom to introduce new classes if some objects do not resemble any labeled observation. We present such an example below, where we consider protein sequences with experimentally confirmed labels (the "true" labels) and others with only machine predicted labels (which we treat as unlabeled objects).

4. Experiments

In a first experiment we compare the proposed TIWD cluster process with several hierarchical clustering methods on synthetic data, generated as follows: (i) a random block-partition matrix B of size $n = 500$ is sampled with $k_B = 10$; (ii) $d = 100$ samples from $N(\mathbf{0}_n, \Sigma)$ are drawn, with $\Sigma = \alpha I_n + \alpha\theta B$, $\alpha = 2$ and different θ-values; (iii) squared Euclidean distances are stored in the matrix $D_{(n\times n)}$.

A two-dimensional kernel PCA projection of an example distance matrix is shown in the left panels of Fig. 2 (large θ ↔ clear cluster separation in the upper panel, and small θ ↔ highly overlapping clusters in the lower panel). 5000 Gibbs sweeps are computed for the TIWD cluster process (after a burn-in phase of 2000 sweeps), followed by an annealing procedure to "freeze" a certain partition, cf. section 3.4. For comparing the performance, several hierarchical clustering methods are applied: "Wards", "complete linkage", "single linkage", "average linkage", see (Jain & Dubes, 1988), and the resulting trees are cut at the same number of clusters as found by TIWD. The right panels show the agreement of the inferred partitions with the true labels, measured in terms of the adjusted rand index. If the clusters are well-separated, all methods perform very well, but for highly overlapping clusters, TIWD shows significant advantages over the hierarchical methods.

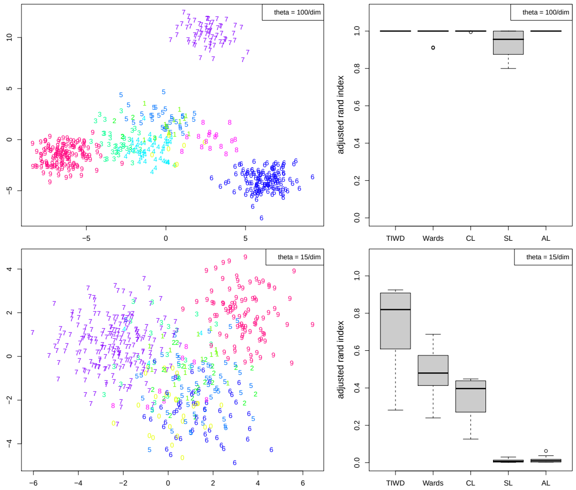

Figure 2. TIWD vs. hierarchical clustering (“Wards”, “complete linkage”, “single linkage”, “average linkage”) on synthetic data ($k = 10$, $n = 500$, $d = 100$, repeated 20 times).

In a second experiment we investigate the scalability of the algorithm. The “small θ”-experiment above (lower panels in Fig. 2) is repeated for $n = 8000$. Figure 3 depicts the trace of the number of blocks k_B during sampling. The sampler stabilizes after roughly 500 sweeps. Note the remarkable stability of the sampler (compared to the usual situations in “traditional” mixture models), which follows from the fact that no label-switching can appear. On a standard computer, this experiment took roughly two hours, which leads us to the conclusion that the proposed sampling algorithm is so efficient (at least for moderate k) that memory constraints are probably more severe than time constraints on standard hardware.

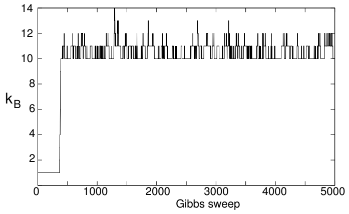

Figure 3. Traceplot of the number of blocks k_B during the Gibbs sweeps for a large synthetic dataset. (10 clusters, $n = 8000$).

In a next experiment we analyze the influence of encoding the translation invariance into the likelihood (our TIWD model) versus the standard WD process and row-mean subtraction as described in section 3.2. A similar random procedure for generating distance matrices is used, but this time we vary the number of replications d and the mean vector $\boldsymbol{\mu}$. If $\boldsymbol{\mu} = \mathbf{0}_n$, both the standard WD process and the TIWD process are expected to perform well, which is confirmed in the 1st and 3rd panel (left and right boxplots). Row-mean subtraction, however, introduces significant bias and variance. For nonzero mean vectors (2nd and 4th panel), standard WD completely fails to detect the cluster structure, and row-mean subtraction can only partially overcome this problem. The TIWD process clearly outperforms the other models for nonzero mean vectors.

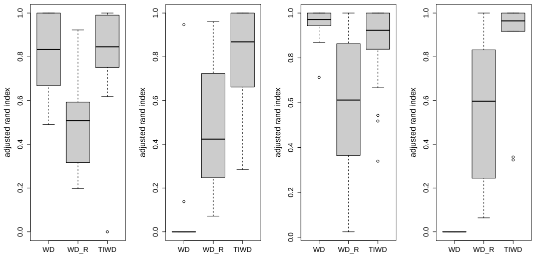

Figure 4. Comparison of WD and TIWD cluster process on synthetic data. ”WD”: standard WD, ”WD_R”: WD with row-mean subtraction. Left to right: (i) $d = 3, \boldsymbol{\mu} = \mathbf{0}$; (ii) $d = 3, \mu_i \sim N(40, 0.1)$; (iii) and (iv): same for $d = 100$.

In a last experiment we consider a semi-supervised application example in which we study all globin-like protein sequences from the UniProtKB database (with experimentally confirmed annotations) and the TrEMBL database (with unconfirmed annotations). The former set consists of 1168 sequences which fall into 114 classes. These sequences form the “supervised” subset, and their assignments to blocks in the partition matrix are “clamped” in the Gibbs sampler. The latter set contains 2603 sequences which are treated as the “unlabeled” observations. Pairwise local string alignment scores s_{ij} are computed between all sequences and transformed into dissimilarities using an exponential transform. The resulting dissimilarity matrix D is not guaranteed to be of negative type (and indeed, $-QDQ$ has some small negative eigenvalues). We overcome this problem by using the randomized low-rank approximation technique according to (Vempala, 2004; Belabbas & Wolfe, 2007), cf. section 3.5, which effectively translates D into a matrix $\tilde{D}$ which is of negative type. The Ewens process model makes it possible to assign the unlabeled objects to existing classes or to new ones. Finally, almost all unlabeled objects are assigned to existing classes, with the exception of three new classes which have an interesting biological interpretation. Two of these classes contain globin-like bacterial sequences from *Actinomycetales*, a very special group of obligate aerobic bacteria which have to cope with oxidative stress. The latter might explain the existence of redox domains in the globin sequences, like the Ferredoxin reductase-type (FAD)-binding domain observed in all sequences in one of the clusters and the additional Nicotinamide adenine dinucleotide (NAD)-binding domain present in all sequences in the second new cluster, see Figure 5. Some of the latter sequences appear to be similar to another class that also contains *Actinomycetales* (see the large "off diagonal" probabilities surrounded by the blue circle) which, however, share a different pattern around some heme binding sites. The third new class contains short sequence fragments which show a certain variant of the hemoglobin beta subunit. With the exception of the above mentioned similarity of one of the Actino-bacterial classes to another one, the three new classes show no similarity to any of the other classes, which nicely demonstrates the advantage of a semi-supervised learning model that is flexible enough to allow the creation of new groups.

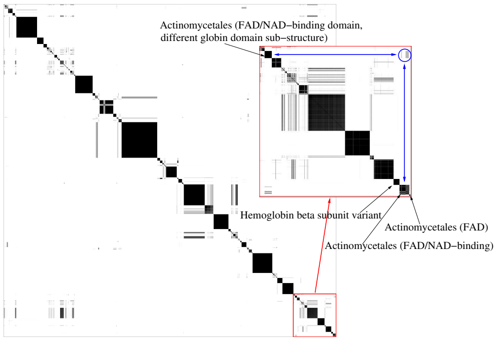

Figure 5. Co-membership probabilities of globin proteins.

5. Conclusion

We introduced a flexible probabilistic model for clustering dissimilarity data. It contains an exchangeable partition process prior which avoids label-switching problems. The likelihood component follows a generalized Wishart model for squared Euclidean distance matrices which is invariant under translations and rotations, under permutations, and under scaling transformations. We call this clustering model the *Translation Invariant Wishart-Dirichlet* (TIWD) cluster process. The main contributions of this work are threefold: (i) On the modelling side, we propose that it is better to work directly on the distances, without computing an explicit dot-product- or vector-space- representation, since such embeddings add unnecessary bias and variance to the inference process. Experiments on simulated data corroborate this proposition by showing that the TIWD model significantly outperforms alternative approaches. In particular if the clusters are only poorly separated, the full probabilistic nature of the TIWD model has clear advantages over hierarchical approaches. (ii) On the algorithmic side we show that costly matrix operations can be avoided by carefully exploiting the inner structure of the likelihood term. We prove that a sweep of a Gibbs sampler can be computed in $O(n^2+nk_B^2)$ time, as opposed to $O(n^4k_B)$ for a naive implementation. Experiments show that these algorithmic improvements make it possible to apply the model to large-scale datasets. (iii) A semi-supervised experiment with globin proteins revealed the strength of our partition process model which is flexible enough to introduce new classes for objects which are dissimilar to any labeled observation. We could identify an interesting class of bacterial sequences, and a subsequent analysis of their domain structure showed that these sequences indeed share some unusual structural elements.

We have implemented a software package for the TIWD model which links efficient `C++` MCMC code to a user-friendly `R`-interface. We will make this package (including the datasets used in this paper) available on `mloss.org`.

Acknowledgement. This work has been partially supported by the FP7 EU project SIMBAD.

References

Anderson, T.W. The non-central wishart distribution and certain problems of multivariate statistics. *Ann. Math. Statist.*, 17(4): 409–431, 1946.

Belabbas, M.A. and Wolfe, P.J. Fast low-rank approximation for covariance matrices. In *IEEE Workshop on Computational Advances in Multi-Sensor Processing*, pp. 293 – 296, 2007.

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

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

Ewens, W.J. The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, 3:87–112, 1972.

Golub, G.H. and Van Loan, C.F. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, MD, USA, 1989.

Jain, A. and Dubes, R. *Algorithms for Clustering Data*. Prentice Hall, 1988.

Jasr, A., Holmes, C.C., and Stephens, D.A. Markov chain monte carlo methods and the label switching problem in Bayesian mixture modeling. *Statist. Sci.*, 20(1):50–67, 2005.

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

McCullagh, P. Marginal likelihood for distance matrices. *Statistica Sinica*, 19:631–649, 2009.

McCullagh, P. and Yang, J. How many clusters? *Bayesian Analysis*, 3:101–120, 2008.

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

Pitman, J. Combinatorial stochastic processes. In Picard, J. (ed.), *Ecole d'Ete de Probabilites de Saint-Flour XXXII-2002*. Springer, 2006.

Srivastava, M.S. Singular Wishart and multivariate beta distributions. *Annals of Statistics*, 31(2):1537–1560, 2003.

Vempala, S. *The Random Projection Method*. Series in Discrete Mathematics and Theoretical Computer Science. AMS, 2004.