Dirichlet Process Mixture Models on Symmetric Positive Definite Matrices for Appearance Clustering in Video Surveillance Applications

Anoop Cherian[*] Vassilios Morellas[†] Nikolaos Papanikolopoulos[‡] Saad J. Bedros[§]

Abstract

Covariance matrices of multivariate data capture feature correlations compactly, and being very robust to noise, they have been used extensively as feature descriptors in many areas in computer vision, like, people appearance tracking, DTI imaging, face recognition, etc. Since these matrices do not adhere to the Euclidean geometry, clustering algorithms using the traditional distance measures cannot be directly extended to them. Prior work in this area has been restricted to using K-means type clustering over the Riemannian space using the Riemannian metric. As the applications scale, it is not practical to assume the number of components in a clustering model, failing any soft-clustering algorithm. In this paper, a novel application of the Dirichlet Process Mixture Model framework is proposed towards unsupervised clustering of symmetric positive definite matrices. We approach the problem by extending the existing K-means type clustering algorithms based on the logdet divergence measure and derive the counterpart of it in a Bayesian framework, which leads to the Wishart-Inverse Wishart conjugate pair. Alternative possibilities based on the matrix Frobenius norm and log-Euclidean measures are also proposed. The models are extensively compared using two real-world datasets against the state-of-the-art algorithms and demonstrate superior performance.

1. Introduction

Clustering of object appearances from video cameras is a long studied problem in computer vision. Appearances of the same object can vary significantly from one camera to the next, for example, in a video surveillance application. This can be due to camera sensor noise, changes in scene illumination, presence of shadows, changes in the pose of the object, partial occlusion, etc. Fig. 1 shows images from a real-world people appearance tracking session and delineates some of these difficulties.

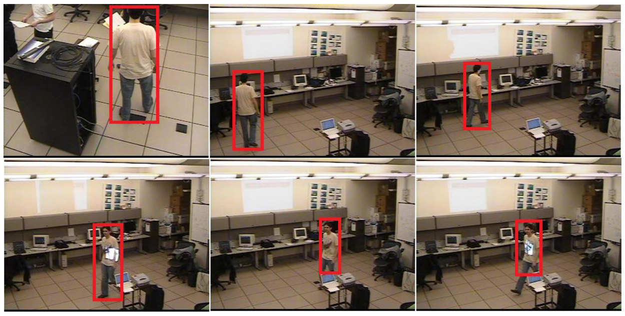


Figure 1: Images from a multi-camera people tracking session. The rectangles drawn on the images mark the person being tracked.

As is clear from Fig. 1, the tracking images can be very noisy and the appearance of the person being tracked can undergo considerable changes. Clustering of appearances in such a setting is a challenging problem. A survey of recent developments in this area can be seen in [13]. It was shown that the covariance matrices of features extracted from the appearances produce significant robustness to the above problems [16]. These matrices provide a non-linear fusion of various image features and have been deployed as good feature descriptors; not only in people tracking, but also in many other areas of computer vision, like face recognition [12], DTI imaging [7], etc. Mathematically, a Covariance Descriptor can be defined as follows:

Definition 1. *Let $F_i \in \mathbb{R}^p$, for $i = 1, 2, \cdots, N$, be the feature vectors from the region of interest of an image, then the Covariance Descriptor of this region $C \in \delta_p^+$ is defined as:*

$$C = \frac{1}{N-1}\sum_{i=1}^{N}(F_i - \mu_F)(F_i - \mu_F)^T \qquad (1)$$

where μ_F is the mean feature vector and δ_p^+ is the space of $p \times p$ Symmetric Positive Definite (SPD) matrices.

A real-time algorithm to compute covariance matrices of object appearances using integral images is proposed in

[*] Department of Computer Science & Engineering, University of Minnesota, Twin Cities. Email: cherian@cs.umn.edu

[†] Department of Computer Science & Engineering, University of Minnesota, Twin Cities. Email: morellas@cs.umn.edu

[‡] Department of Computer Science & Engineering, University of Minnesota, Twin Cities. Email: npapas@cs.umn.edu

[§] Honeywell, MN, USA. Email: saad.bedros@honeywell.com

[16], thus making these descriptors an ideal platform for people tracking in video surveillance applications. On the negative side, covariances do not conform to the Euclidean geometry; but spans the cone of SPD matrices and form a connected Riemannian manifold with an associated Riemannian metric [10]. The mean of a set of covariance matrices can be computed using the Karcher mean algorithm [17], thus suggesting the viability of a K-means type clustering [15] scheme. In [23], a semi-supervised learning framework is suggested by embedding these matrices into the Euclidean space by vectorizing the symmetric matrices created through a log-Euclidean mapping. An EM algorithm for clustering SPD matrices using a mixture of Wishart distributions is suggested in [11]. Since the number of components in the model need to be specified in any soft-clustering algorithm, these models are not scalable to a real-world scenario (for example, clustering appearances of people in a camera surveillance network), motivating investigations toward unsupervised models in which the number of clusters is dynamically updated monotonically according to the complexity of an ever increasing volume of data.

To this end, this paper presents a novel application of the Dirichlet Process Mixture Model (DPMM) framework for clustering SPD matrices. DPMMs provide a standard technique for unsupervised Bayesian clustering and has been successfully utilized in a variety of domains like genomics [28], vision [25], data modeling [4], etc. All these references use a Gaussian-Inverse-Wishart (GIW) DPMM for clustering vector valued data. A multivariate Gaussian distribution is assumed on the data; the cluster mean of which being sampled from a prior Gaussian model, and the covariance matrix being sampled from an inverse-Wishart (IW) distribution. In [26], a translation invariant Wishart DPMM is presented for clustering distance data. This is similar to the GIW model, except that the cluster covariance matrix comes from a Wishart distribution. Bayesian density estimation for functional data analysis is proposed in [21]. To the best of our knowledge, the Dirichlet process framework has never been applied to clustering SPD matrices before.

The basic problem to be tackled in developing a mixture model is to define a probability measure on the underlying data that captures its structure effectively. Some of the well-known statistical measures on SPD matrices are the matrix Frobenius norm, log-Euclidean distance and the logdet divergence. The first two measures can be used to embed the data matrices into the Euclidean space, while the third measure operates directly in the matrix space. Since a Euclidean embedding can essentially distort the structure of the data, we primarily focus in developing the framework in this paper based on the logdet divergence measure, and systematically derive the associated model components toward Bayesian clustering.

The paper is organized as follows: We start with a review of DPMMs in Section 2, followed by a review of the background in Section 3; various DPMMs are introduced in Section 4 that precedes experiments and results in Section 5, finally concluding in Section 6.

2. Dirichlet Process Mixture Model

Nonparametric Bayesian techniques seek a predictive model for the data such that the complexity and accuracy of the model grows with the data size. The existence of such a statistical model is invariably dependent on the property of exchangeability [19] of observations, leading to the De Finetti's theorem, which states that if a sequence of observations $y_1, y_2, \cdots, y_n$ is infinitely exchangeable, then there exists a mixture representation for the joint distribution of these observations. That is,

$$p(y_1, y_2, \cdots, y_n) = \int_\Theta p(\theta) \prod_{i=1}^{n} p(y_i|\theta) d\theta \tag{2}$$

where Θ is an infinite-dimensional space of probability measures and $d\theta$ defines a probability measure over distributions.

A Dirichlet Process, $DP(\alpha, H)$, parameterized by a concentration α and a prior H, is a stochastic process that defines a distribution over probability distributions adhering to Eq. (2) such that if $A_1, A_2, \cdots, A_r$ represent any finite measurable partition of Θ, and if $G \sim DP(\alpha, H)$, then the vector of joint distributions of samples from G over these partitions follow a Dirichlet distribution, $Dir(.)$ [9]. That is,

$$(G(A_1), \cdots, G(A_r)) \sim Dir\left(\alpha H(A_1), \cdots, \alpha H(A_r)\right) \tag{3}$$

In our pertinent problem of finding the number of clusters in the given dataset, we would like to find the distribution G over each of the clusters automatically, by computing the posterior distribution of G given the observations and the prior model H. Fortunately, as is shown in [9], the posterior distribution has the following simple form:

$$p\left(G|y_1, \cdots, y_n\right) \sim Dir\left(\alpha H(A_1) + n_1, \cdots, \alpha H(A_n) + n_r\right)$$

$$\sim DP\left(\alpha + n, \frac{1}{\alpha + n}\left(\alpha H + \sum_{i=1}^{n} \delta_{y_i}\right)\right) \tag{4}$$

where $n_1, n_2, \cdots, n_r$ represent the number of observations falling in each of the partitions $A_1, A_2, \cdots, A_r$ respectively, n is the total number of observations, and δ_{y_i} represents the delta function at the sample point y_i. There are some subtle points to be noted from the Eq. (4): (i) DP acts as a conjugate prior for the distribution over distributions, and (ii) each observation only updates the corresponding component in the Dirichlet distribution. The latter property implies that the underlying distribution is essentially discrete with probability one and is agnostic over the topology

1) Define a likelihood distribution F on data using a suitable distance measure.
2) Find a prior H that is conjugate to F.
3) Model a collapsed Gibbs sampler over the posterior distribution from (1) and (2) as follows:
   3a) Remove a data point y from a cluster C and update the sufficient statistics of C.
   3b) Compute the predictive distribution $p(y|y_{C_i})$ for each of the existing clusters; y_{C_i} represents data in cluster C_i. Also, compute the probability of formation of a new cluster defined by the concentration parameter α of the DP model.
   3c) Sample a cluster from $p(y|y_{C_i})$ and assign the data point to that cluster.
   3d) Repeat the steps (3a), (3b) and (3c) until convergence.

Table 1: Overview of DPMM algorithm

of the underlying space in which the data lie. It was shown in [22] that this property leads to a concentration of the posterior distribution towards the mean, leading to a clustering effect [24].

For the above model to be practically useful and tractable, it is a general practice to model the dependency of the observations y_j to G through a parameterized family $F(\theta_i)$ (where F is the likelihood function with parameters θ_i), leading to a mixture model characterization of the DP as follows:

$$\begin{aligned} y_i|\theta_i &\sim F(\theta_i) \\ \theta_i|G &\sim G \\ G &\sim DP(\alpha, H) \end{aligned} \tag{5}$$

Since the exact computation of the posterior is infeasible when data size is large, we resort to a variant of MCMC algorithms, namely, the collapsed Gibbs sampler [14] for faster approximate inference.

The discussion in this section and the formulas we seek in the context of covariance matrices are summarized in Table 1.

3. Mathematical Preliminaries

This paper approaches the problem of deriving the DPMMs for covariance clustering as an extension to the hard-clustering algorithms. Thus we find analogues of three traditional distance measures in the Bayesian context, namely (i) logdet divergence (ii) matrix Frobenius norm, and (iii) log-Euclidean embedding. This section details the mathematical preliminaries required for our analysis in the subsequent sections.

3.1. LogDet Divergence

The LogDet divergence, D_{ld}, (also known as *Stein's loss*) defines the KL-divergence between two equal-mean Gaussian distributions [5]. Given two SPD matrices $C_1, C_2 \in \delta_p^+$, we have:

$$D_{ld}(C_1, C_2) = Tr(C_1C_2^{-1}) - log|C_1C_2^{-1}| - p \tag{6}$$

where $| \,.\, |$ stands for the matrix determinant. This divergence is not a metric, since it is not symmetric and does not satisfy the triangle inequality. Nevertheless, it is a Bregman matrix divergence for the convex function $-log|.|$, and has been utilized in soft clustering algorithms [6].

3.2. Matrix Frobenius Norm

Given $C_1, C_2 \in \delta_p^+$, the matrix Frobenius distance $D_F(C_1, C_2)$ measures the Euclidean distance between each of the elements of the matrices and is defined as:

$$D_F(C_1, C_2) = \|C_1 - C_2\|_F = \|Vec(C_1) - Vec(C_2)\|_2 \tag{7}$$

where $Vec(.)$ is the matrix *vectorization* operator. It is easy to show that $D_F^2(.,.)$ is a Bregman divergence.

3.3. Log-Euclidean distance

Given $C \in \delta_p^+$, the matrix logarithm $log(C)$, is a symmetric matrix and is no more restricted to the cone of SPD matrices. Using this observation, [1] proposes the log-Euclidean Riemannain metric $D_{le}(C_1, C_2)$ as follows:

$$D_{le}(C_1, C_2) = \|\log(C_1) - \log(C_2)\|_F \tag{8}$$

where $C_1, C_2 \in \delta_p^+$. Note that this metric is similar to Eq. (7), except that it first projects the matrices into their tangent space using the log operator and later embed them into the Euclidean space, which might reduce the distortion produced otherwise from a direct embedding.

3.4. Wishart Distribution

Let $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N$, ($\mathbf{x}_i \in R^p$), be iid random vectors such that $\mathbf{x}_i \sim \mathcal{N}(0, \Sigma)$, for $i = 1, 2, \cdots, N$ and let $X \in \delta_p^+$ such that $X = \sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T$. If we define $n = N - 1$, then X is said to follow a non-singular p-dimensional Wishart distribution $W(n, p, \Sigma)$, with n degrees of freedom ($n > p - 1$) and scale matrix Σ if it has a probability density defined by:

$$W(X; n, p, \Sigma) = \omega(n, p)\frac{|X|^{n-p-1/2}}{|\Sigma|^{n/2}} exp\left\{-\frac{1}{2}tr(\Sigma^{-1}X)\right\}, \tag{9}$$

where $\omega(n, p)$ is a normalizing constant [8].

4. DPMM on SPD Matrices

In this section, we derive the probability densities and the predictive distributions associated with the base measures introduced above.

4.1. LogDet-Wishart Connection

The logdet divergence is a Bregman matrix divergence and thus utilizing the Bregman-exponential family bijection [2], we can derive the associated likelihood distribution. It turns out that this exponential family is the Wishart distribution as stated in the following theorem:

Theorem 1. *Let X_c be the covariance matrix of N iid zero-mean Gaussian-distributed random vectors $\mathbf{x}_i$, i.e. $X_c = \frac{1}{n}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T$ where $\mathbf{x}_i \in \mathbb{R}^p$, $\mathbf{x}_i \sim \mathcal{N}(0,\Sigma)$ and $n = N-1$. Then the probability density of X_c follows:*

$$p(X_c|N,p,\Sigma) = W(X_c;n,p,\Sigma) \propto \exp\left\{-\frac{1}{2}D_\psi(\Sigma,X_c)\right\} p_0(X_c),$$

where D_ψ is the Bregman matrix divergence function for the convex function $\psi(X) = -n\ log|X|$ and p_0 is the base measure.

Proof. Let $X \in \delta_p^+$ and assume $X = \sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T$. Thus, $X = nX_c$. From the definition of the Wishart distribution, we have $X \sim W(n,p,\Sigma)$. Substituting for X and rearranging the terms, we get:

$$p(X|N,p,\Sigma) \propto |X_c|^{-(p+1)/2}\exp\left\{-\frac{n}{2}\left[Tr\left(\Sigma^{-1}X_c\right) - log|\Sigma^{-1}X_c|\right]\right\}$$
$$\propto \exp\left\{-D_\psi(\Sigma,X_c)\right\}|X_c|^{-(p+1)/2}$$

□

It is well-known in multivariate statistics that for the Wishart distribution $W(n,p,\Sigma)$, the conjugate prior is the Inverse-Wishart distribution parametrized as $IW(n,p,S)$ where S is the inverse scale matrix. Utilizing this observation, we derive the associated predictive distribution for the Wishart-Inverse-Wishart (WIW) DPMM in the next subsection.

4.2. Predictive Distribution

As was mentioned in the algorithm described in Table 1, the next step to build the DPMM is to formulate a collapsed Gibbs sampling framework for inference, which requires deriving the predictive distribution. This is given in the following theorem.

Theorem 2. *Let $X_i \in \delta_p^+$, $i = 1,2,\cdots,N-1$, belong to a cluster C such that $X_i \sim W(n,p,\Sigma)$, where Σ is the Wishart scale matrix and n, the degrees of freedom. Let $\Sigma \sim IW(n,p,S)$ with inverse scale matrix S. Then the predictive distribution of a data matrix X_N to belong to the cluster $C = \{X_1,X_2,\cdots,X_{N-1}\}$ will be:*

$$p(X_N|X_1X_2...X_{N-1}) = \int_{\delta_p^+} p(X_N|\Sigma)\ p(\Sigma|X_1X_2...X_{N-1})d\Sigma$$
$$= \frac{\omega\left((N+1)n,p\right)}{\omega(n,p)\ \omega(Nn,p)}\ \frac{|X_N|^{\frac{(n-p-1)}{2}}\ |\sum_{i=1}^{N-1}X_i + S|^{\frac{Nn}{2}}}{|\sum_{i=1}^{N}X_i\ +\ S|^{\frac{(N+1)n}{2}}}$$

where $\omega(n,p) = \int_{\delta_p^+}|Y|^{\frac{n-p-1}{2}}\exp\left\{-\frac{1}{2}tr(Y)\right\}dY$.

Proof. See Appendix A. □

4.3. Frobenius Norm based DPMM

Using the base measure as the matrix Frobenius norm, it can be shown that the exponential family for the associated Bregman divergence is the multivariate normal distribution. That is, given X, $\mu_x \in \delta_p^+$, and a variance σ^2,

$$p(X|\mu_X,\sigma^2) \propto \exp\left(-\frac{\|X-\mu_X\|_F^2}{2\sigma^2}\right) \propto \exp\left(-\frac{\|\mathcal{V}(X)-\mathcal{V}(\mu_X)\|_2^2}{2\sigma^2}\right) \quad (10)$$

where $\mathcal{V} : \delta_p^+ \rightarrow \mathbb{R}^{p(p+1)/2}$ is the half-vectorization operator. The variance σ^2 in the Eq. (10) can be generalized using a covariance matrix Σ between the vectorized matrices, leading to a standard GIW DPMM, where the $\mathcal{V}(\mu_X) \sim \mathcal{N}(\mu,S_1)$ and $\Sigma \sim IW(n,p,S_2)$; S_1,S_2 being the hyper-scale matrices and μ is the hyper-mean [3].

4.4. Log-Euclidean based DPMM

Similar to the approach above, we can derive the associated density function for the log-Euclidean distance, Eq. (8). Using the Euclidean embedding suggested in [17], the density function takes the form:

$$p(X|\mu_X,\sigma^2) \propto \exp\left(-\frac{1}{2}\frac{\|\mathcal{V}(log(X))-\mu_x\|_2^2}{\sigma^2}\right) \quad (11)$$

where $\log(.)$ is the matrix logarithm, $\mu_X = \mathcal{V}(1/N\sum_{i=1}^{N}\log(X_i))$ and σ^2 is the assumed variance. We can approximate μ_X to follow a multivariate normal distribution, in which case the DPMM follows the standard GIW model as mentioned earlier.

5. Experiments and Results

This section details the empirical evaluation of the DPMMs to the state-of-the-art clustering techniques on covariance matrices. First, we introduce two metrics on which the performance is defined. Later, results on simulated data and real appearance data are presented.

5.1. Purity

A standard way to compare clustering performance is the rand-index measure [20]. However, to better explore the representation power of covariance matrices as a means of data representation and to evaluate the ability of the DPMM to automatically cluster the data, we define variants of the rand-index measure which we call *purity*. Performance results of our methodology are analyzed and presented in the light of two such purity measures: (i) cluster purity and, (ii) class purity. Cluster purity captures the ability of the proposed methodology (and the associated metric employed) to partition the symmetric positive definite matrix data in the multi-dimensional space they exist. It is defined for every cluster automatically created by the DPMM process as the fraction of class instances that dominate the respective cluster. For example, in Fig. 2, cluster 1 is dominated by class instances denoted by the triangles although instances belonging to another class (denoted by the stars) are also included in the same cluster. Mathematically, we can write this measure as follows: for a cluster C_i, if $label(C_i)$ represents the set of labels of all the data points in C_i, then we define the cluster Purity, $P_{cluster}$, of C_i as:

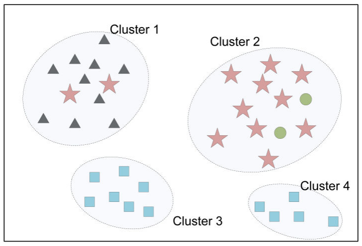


Figure 2: Illustration of the measure of purity

$$P_{cluster}(i) = \frac{\#\{label(C_i) = \ell^*\}}{\#C_i} \text{ where } \ell^* = \max_{\ell} \#\{label(C_i) = \ell\}, \tag{12}$$

where #{.} defines the cardinality of the set. This metric alone might not be a good measure of the clustering performance. For example, in the cluster 2 in Fig. 2, there are circle labels which have a low cardinality as compared to star labels and are ignored in $P_{cluster}$. Another problem being when data gets split into multiple clusters, each cluster being pure in itself (for example, cluster 3 and cluster 4 in Fig. 2). To account for this, we use the class purity P_{class} metric. Class purity captures the complexity of the data exposed by the form of data representation used (in our case covariance matrices of intensity and gradient intensity information). It also reflects the clustering challenge experienced by the DPMM whereby instances of a single class are assigned to multiple clusters not necessarily dominating the assigned cluster. Thus class purity helps better understand if the feature vectors that we chose for building the covariances adequately capture the differentiating properties of the classes. For a label ℓ,

$$P_{class}(\ell) = \frac{\#\{label(C_{k^*}) = \ell\}}{\#C_{k^*}} \text{ where } k^* = \max_{k} \#\{label(C_k) = \ell\} \tag{13}$$

The *Purity*, P, of clustering is defined as the weighted sum of Eq. (12) and Eq. (13) for all the clusters, i.e.

$$P = \frac{1}{2}\left[\sum_{i=1}^{i=K} \frac{P_{cluster}(i)}{K} + \sum_{\ell=1}^{\ell=L} \frac{P_{class}(\ell)}{L}\right] \tag{14}$$

for K clusters produced by the algorithm and L true class labels. For a perfect clustering, purity will be unity and the number of clusters discovered by the DPMM should match the true number.

5.2. Simulated Experiment

This section evaluates the correctness of the WIW DPMM on a small simulated dataset. The dataset consisted of 100 covariance matrices of dimension 5 × 5. Each covariance matrix was generated from 1K normal distributed random vectors of dimension 5×1, thus fixing the degrees of freedom parameter in the WIW model. Six different Gaussian distributions were used to create the dataset. Fig. 3 plots an ISOMAP embedding of the clusters found by the DPMM after 10 iterations of the Gibbs sampling. The purity for this clustering was found to be 0.97.

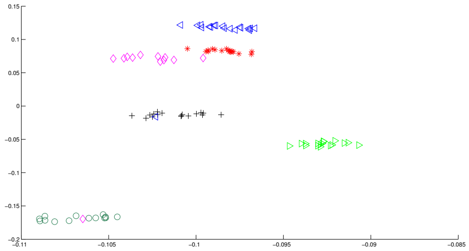


Figure 3: ISOMAP embedding of clustered covariances on 100 covariance matrices with six true clusters. Data belonging to each cluster is shown with a unique symbol.

Increasing Cluster Sizes The accuracy of clustering for an increasing number of true clusters is evaluated in this section for the WIW model. The true clusters were increased from 10 to 100, with each of the true clusters having atleast five covariance matrices from the same underlying normal distribution. Fig. 4 shows the results of this experiment. As is clear from the plot, the model was able to find the clusters with an average purity score of more than 0.8, while maintaining the number of clusters discovered close to the ground truth.

5.3. Experiments on Real-Data

This section details the datasets used for performance evaluation. The choice of the datasets were based on three properties of the DPMM that we thought to evaluate: (i) performance on noisy real-world data, (ii) robustness to clustering appearances from multiple cameras, and (iii) performance against the size of covariances. For (i) and (ii), the dataset shown in Fig. 5 was used, which contained background subtracted people appearances from real-time tracking sessions in our lab recorded using two cameras. The dataset contained 900 covariances and 30 true clusters. We used a five dimensional feature vector,

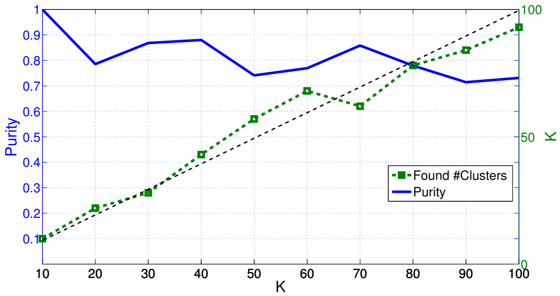


Figure 4: Simulations with the true number of clusters increasing from 10 to 100. Left y-axis is Purity, right y-axis is the number of clusters discovered and x-axis shows the ground truth. The dotted straight line is drawn for an ideal case comparison.

$F = [I_R, I_G, I_B, I_x, I_y]^T$, to construct each covariance matrix, where the first three dimensions of F are the pixel intensities in the three color channels red, green and blue respectively, while I_x, I_y stand for the gray level pixel gradients in x and y directions respectively. Thus each covariance matrix used in this experiment was 5×5. All the appearances were resized to 100×100 and 3000 points were randomly sampled, thus fixing the number of degrees of freedom in the DPMM. The true cardinality of the clusters varied from 5 to 50. To evaluate (iii), we used the FERET face appearance dataset [18]; a few sample images of which are shown in Fig. 5. We used 110 different face classes from this dataset, with each class containing 7 different poses of the same person. Covariances of size 40×40 were created using 40 Gabor-filters as suggested in [12].

Figure 5: Top: Sample images from the appearance tracking dataset. Bottom: Sample images from the FERET face appearance dataset.

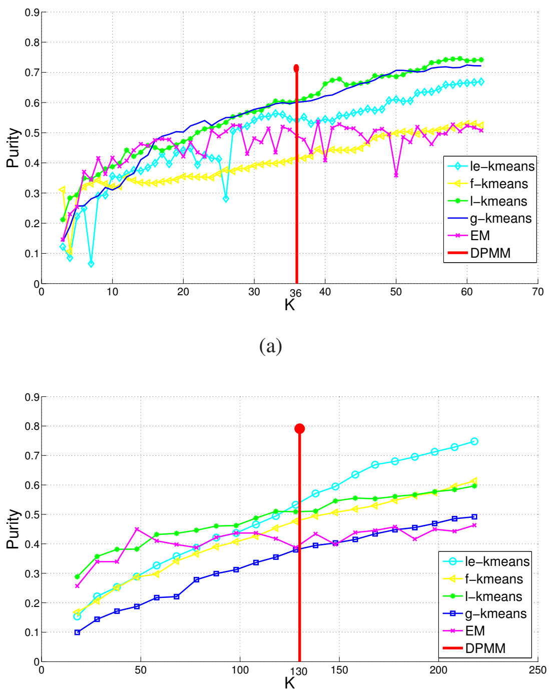
5.4. Setup

The algorithms were implemented in Matlab. The concentration parameter of the DPMM was initialized to 1 and then re-estimated at each iteration from a mixture of gamma distributions as described in [27]. A gamma prior, G(30,60), gave the best performance for all the three DPMMs. Hyperparameters of the models were estimated by taking the mean of all the covariances in the dataset, while the initial allocation of the data points was done using K-means algorithm. The collapsed Gibbs sampler for the inference model converged in less than 20 iterations.

5.5. Results

The clustering performance of the WIW model for the two datasets is given in Fig.6(a) and Fig. 6(b). The model is compared against K-means algorithms using the Riemannian geodesic distance (g-kmeans), the symmetric logdet divergence (l-kmeans), Matrix Frobenius distance (f-kmeans), log-Euclidean distance (le-kmeans) and the EM-algorithm based on mixture of Wishart distributions [11]. The cluster means for the K-means was computed using the Karcher mean algorithm described in [17]. Since we assume we do not know the true number of components in the hard/soft-clustering algorithms, the number of components in each model was increased from 3 to 60 for the people dataset and from 18 to 220 for the face dataset; which forms the x-axis of the plots. As is clear from the plots, the WIW model automatically figures out the true clusters and its purity score was found to outperform the scores of all the other algorithms, at the same time keeping the identified number of clusters close to the true number. This dominance of the WIW model is more discernible in the plots from the face dataset Fig. 6(b).

Next, we evaluate the performance of the DPMMs against each other on the above datasets. Fig. 7 shows the results on the people appearances dataset for the Frobenius-DPMM (P-Frob), the log-Euclidean-DPMM (P-LE) and the WIW model (P-WIW), along with the results on the face appearance dataset (F-Frob, F-LE and F-WIW respectively). As is clear from the plots the WIW model performs better than the vectorization based methods. This argument goes stronger with the face dataset, where vectorization methods did not seem applicable at all, perhaps the reason being the curse of dimensionality, as the vectorization of a 40×40 matrix produces an 820 dimensional vector, which requires a very large dataset for effective clustering. This seems to be a major limitation with the vectorization methods. In addition, we also observed that for all the three models, the number of clusters discovered in the people appearance data remained in the range of 24 to 36 (30 being the ground truth), while the F-WIW model found 130 clusters (110 being ground truth); those for the vectorization methods of the face dataset deviated considerably from the true number. Another observation to be pointed out from the class purity scores from Fig. 7 is their high value for the faces dataset, while relatively lower scores for the people appearances. This points to the inadequacy of the features used in creating the covariances for the latter dataset and the scope for improvement. A sample visualization of clustering us-



(a)

(b)

Figure 6: (a) compares the clustering performance using the People dataset and, (b) compares the clustering performance using the FERET face dataset. x-axis is the number of components used for the hard/soft clustering algorithms and y-axis measures purity.

ing the WIW model for a random subset of the people appearances dataset (115 appearances and 3 true clusters) is shown in Fig. 8.

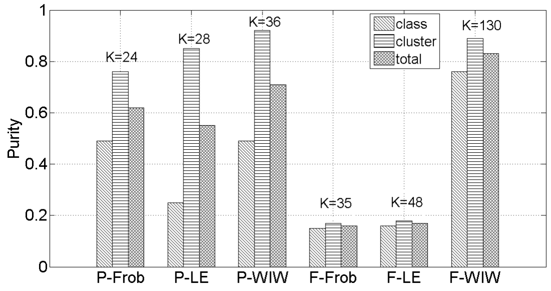


Figure 7: Comparison of various DPMMs for people appearance and face datasets. The K value in the x-axis shows the number of clusters found by the DPMM; the ground truth being K=30 for the first three plots and K=110 for the last three.

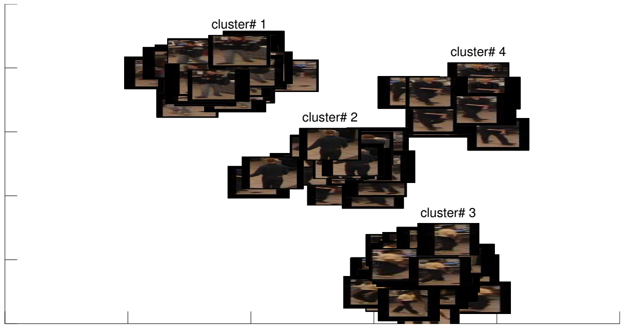


Figure 8: A visualization of clustering using WIW model on a subset of the people appearances dataset. The result is for a purity score of 0.83, while the true number of clusters is 3 (Note that appearances in clusters 2 and 4 are actually of the same person, but differs slightly).

Dataset	*EM*	*K − means*	*DPMM*
People(#clusters)	1.49(50)	2.3(50)	**3.48(50)**
Face(#clusters)	160.11(10)	116.54(10)	**11.26(50)**

Table 2: Computational performance of various clustering algorithms. Each entry in the table lists the time taken (in seconds) for one iteration of the algorithm. Number of clusters is shown inside bracket.

Computational Cost: Table. 2 compares the cost per iteration of various algorithms. Since the complexity is proportional to the number of clusters the algorithm operates on, this value is also shown. WIW model seems to scale well with the dimensionality of the data compared to other algorithms. This is justified as there is no closed form solution to compute the mean of the covariances as in the K-means case and the need to compute matrix inverse as in the case of EM with mixture of Wishart distribution.

6. Conclusion and Future Work

In this paper, an unsupervised clustering framework was introduced over the space of SPD matrices using DPMMs. We investigated three different models based on the Frobenius distance, the log-Euclidean and the Wishart-Inverse-Wishart model. The results clearly demonstrated the effectiveness of the models. The experiments also expounded the superiority of the Wishart model over the vectorization approaches when the feature dimensions were large. Going forward, a direction to investigate is in adapting the models to hierarchical Dirichlet processes.

7. Acknowledgements

The authors would like to thank Prof. Arindam Banerjee (University of Minnesota) for many helpful discussions. This material is based upon work supported in part by the

U.S. Army Research Laboratory and the U.S. Army Research Office under contract #911NF-08-1-0463 (Proposal 55111-CI), Honeywell, and the National Science Foundation through grants #IIP-0443945, #CNS-0821474, #IIP-0934327, #CNS-1039741, and #SMA-1028076.

A. Marginal Distribution

This section derives marginal distribution of a Wishart distributed data matrix X given the hyperparameter S over the space of all canonical parameter matrices Σ.

Suppose that $X \in \delta_p^+$ and $X \sim W(n, p, \Sigma)$ and let $\Sigma \sim IW(n, p, S)$. Then, $P(X|S) = \int_{\delta_p^+} W(X; n, p, \Sigma)\, IW(\Sigma; n, p, S) d\Sigma$. Using the Jacobian for $(\Sigma^{-1} \Rightarrow \Sigma)$, we have $d\Sigma = \frac{dR}{|R|^{-(p+1)}}$, where $R = \Sigma^{-1}$, which can then be used to rewrite $P(X|S)$ as

$$\Rightarrow \frac{1}{c(r,p)^2} |S|^{\frac{r}{2}} |X|^{\frac{r-p-1}{2}} \int_{\delta_p^+} |R|^{\frac{2r-p-1}{2}} exp\left[-\frac{1}{2} tr(R(X+S))\right] dR \tag{15}$$

Now, let $Y = R(X+S)$. Then $dY = |X+S|^{\frac{p+1}{2}} dR$. Substituting this in Eq. (15) and rearranging the terms we have the desired result.

References

[1] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *Matrix Analysis and Applications*, 29(1):328, 2008. 3419

[2] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005. 3420

[3] M. Beal, Z. Ghahramani, and C. Rasmussen. The infinite hidden Markov model. *Adv. in Neural Info. Processing Systems*, 1:577–584, 2002. 3420

[4] N. Bouguila and D. Ziou. A dirichlet process mixture of generalized dirichlet distributions for proportional data modeling. *Neural Networks*, 21(1):107 –122, jan. 2010. 3418

[5] J. Davis and I. Dhillon. Differential entropic clustering of multivariate gaussians. *Adv. in Neural Info. Processing Systems*, 19:337, 2007. 3419

[6] J. Davis, B. Kulis, P. Jain, S. Sra, and I. Dhillon. Info.-theoretic metric learning, 2007. 3419

[7] I. Dryden, A. Koloydenko, and D. Zhou. Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *University of Nottingham, NG7 2RD, UK*, 2008. 3417

[8] M. Eaton. *Multivariate statistics: a vector space approach*. Wiley New York, 1983. 3419

[9] T. Ferguson. A Bayesian analysis of some nonparametric problems. *The annals of statistics*, 1(2):209–230, 1973. 3418

[10] W. Forstner and B. Moonen. A metric for covariance matrices. *Qua vadis geodesia*, pages 113–128, 1999. 3418

[11] S. Hidot and C. Jean. An expectation-maximization algorithm for the wishart mixture model: Application to movement clustering. *Patt. Reco. Lett.*, 31:2318–2324, 2010. 3418, 3422

[12] C. Liu. Gabor-based kernel PCA with fractional power polynomial models for face recognition. *Pattern Analysis and Machine Intelligence*, 26(5):572–581, 2004. 3417, 3422

[13] T. Moeslund, A. Hilton, and V. Kruger. A survey of advances in vision-based human motion capture and analysis. *Comp. Vis. and Image Understanding*, 104(2-3):90–126, 2006. 3417

[14] N. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, pages 249–265, 2000. 3419

[15] O. Tuzel, F.Porikli, and P. Meer. Covariance Tracking using Model Update Based on Lie Algebra. *Comp. Vis. and Patt. Reco., Vol. 1, pp. 728–735*, June 2006. 3418

[16] O. Tuzel, F.Porikli, and P. Meer. Region Covariance: A Fast Descriptor for Detection and Classification. *Euro. conf. on Computer Vision*, May 2006, TR2005–111. 3417, 3418

[17] X. Pennec, P. Fillard, and N. Ayache. A Riemannian framework for tensor computing. *Intl. Journal of Computer Vision*, 66(1):41–66, 2006. 3418, 3420, 3422

[18] P. Phillips, H. Moon, S. Rizvi, and P. Rauss. The FERET evaluation methodology for face-recognition algorithms. *Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000. 3422

[19] J. Pitman and J. Picard. *Combinatorial stochastic processes*. Springer, Berlin, 2006. 3418

[20] W. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971. 3420

[21] A. Rodriguez, D. Dunson, and A. Gelfand. Bayesian nonparametric functional data analysis through density estimation. *Biometrika*, 96(1):149, 2009. 3418

[22] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639–650, 1994. 3419

[23] R. Sivalingam, V. Morellas, D. Boley, and N. Papanikolopoulos. Metric Learning for Semi-supervised Clustering of Region Covariance Descriptors. *Intl. Conf. on Distr. Smart Cameras (ICDSC)*, 6:1–8, 2009. 3418

[24] E. Sudderth. *Graphical models for visual object recognition and tracking*. PhD thesis, 2006. 3419

[25] E. Sudderth, A. Torralba, W. Freeman, and A. Willsky. Describing visual scenes using transformed dirichlet processes. *Adv. in neural Info. processing systems*, 18:1297, 2006. 3418

[26] J. E. Vogt, S. Prabhakaran, T. J. Fuchs, and V. Roth. The translation-invariant Wishart-Dirichlet process for clustering distance data. *Intl. conf. on Machine Learning*, pages 1111–1118, June 2010. 3418

[27] M. West. Hyperparameter estimation in Dirichlet process mixture models. *Inst. of stat. and decision sciences*, 1992. 3422

[28] E. Xing, M. Jordan, and R. Sharan. Bayesian haplotype inference via the Dirichlet process. *Journal of Computational Biology*, 14(3):267–284, 2007. 3418