MODELS WITH PRODUCTS OF DIRICHLET PROCESSES

Petar M. Djurić and André Ferrari***

*Department of Electrical & Computer Engineering
Stony Brook University, USA

**Université de Nice-Sophia Antipolis, CNRS
Observatoire de la Côte d'Azur, 06108 Nice cedex, France

email: petar.djuric@stonybrook.edu, ferrari@unice.fr

ABSTRACT

Nonparametric Bayesian models are often preferred over parametric models due to their superior flexibility in interpreting data. A strong motivation for the use of these models is the desire of avoiding the assumptions that are necessary for parametric models. A prominent place in Bayesian nonparametrics is played by the Dirichlet process, which is defined by a base measure and a concentration parameter. In this paper, we propose the construction of models based on products of Dirichlet processes and corresponding mixture models. We show how these processes can be used for classification of data with shared features. The proposed processes are different from the recently introduced hierarchical Dirichlet processes. We show the use of the proposed model on classification of multivariate time series and demonstrate its performance with computer simulations.

Index Terms— Dirichlet processes, collapsed Gibbs sampling, Dirichlet mixture models

1. INTRODUCTION

Nonparametric Bayesian methods are steadily gaining interest in the machine learning community [1, 2, 3]. A central place in nonparametric Bayesian theory is occupied by the Dirichlet processes (DPs) [4, 5]. A DP is defined by a base measure (or base distribution) H and a concentration parameter $\alpha > 0$, and it produces draws of discrete distributions with probability one. In applications, where we have to classify data, we use DP mixture (DPM) models [6, 7, 8], where the classes and their parameters are generated from DP(α, H), and the data for parametric distributions given the drawn parameters from a distribution F [9]. The mixture model in principle may be with a countably infinite number of clusters. A typical implementation of classification is carried out by Markov chain Monte Carlo sampling [10].

The classification with DPM models proceeds in a non-supervised fashion, and it does not require information about the number of classes of the data [5]. In theory, this number can grow to infinity as the number of data for classification grows to infinity. With DPMs, classification is combined with the task of determining the number of classes. If there are new data, one does not have to restart the classification process, and instead, the new data either join one of the existing classes or they form a new class.

A relatively recent development in nonparametric Bayesian methods has been the introduction of hierarchical DPs [11]. They were developed with the intent to model groups of data, where observations from a group are draws from a mixture model and where the groups share mixture components. More precisely, if G_0 is a global probability measure drawn from DP(α, H), one then defines $G_j|\alpha_0, G_0$ to be generated from DP(α_0, G_0). For such a hierarchical DP, one can construct a hierarchical DPM model.

In this paper, we propose a completely different construction of models sharing mixture components than that followed by hierarchical DPs. It is based on a product of DPs (PDPs) and with it we can readily obtain PDP mixture models. We show how these models are used for classifying multivariate time series and demonstrate their performance with simulations.

The structure of the paper is as follows. In the next section, we review the standard Dirichlet mixture models. In Section 3, we introduce the notion of PDPs. An example that shows how we use the novel PDP mixture model for classification is provided in Section 4. Simulation results of classification are presented in Section 5 and conclusions in Section 6.

2. STANDARD DIRICHLET MIXTURE MODELS

We first recall the definition of a DP. Suppose H is a probability measure over $(\Omega, \mathbb{F})$, where $(\Omega, \mathbb{F})$ is some measurable space.

Definition [4]: *The random probability measure G defined over $(\Omega, \mathbb{F})$ is distributed according to the DP(α, H), where $\alpha > 0$, if for any finite partition $A_1, A_2, \cdots, A_k$, of Ω, $\big(G(A_1), G(A_2), \cdots, G(A_k)\big)$ is distributed according to the Dirichlet distribution defined by*

$$\big(G(A_1), \cdots, G(A_k)\big) \sim \mathrm{Dir}\big(\alpha H(A_1), \cdots, \alpha H(A_k)\big). \quad (1)$$

In (1), the probability measure H is referred to as base measure and α as concentration parameter.

We can obtain a realization of a DP(α, H) by the following scheme: If in the previous $n-1$ samples, the different number of labels is L,

1. Generate labels according to

$$z_n|\{z_1, z_2, \cdots, z_{n-1}\} \sim \begin{cases} p_l = \frac{n_l}{\alpha+n-1}, & l = 1, 2 \cdots, L \\ p_{L+1} = \frac{\alpha}{\alpha+n-1}, & l = L+1, \end{cases} \quad (2)$$

   where n_l is the number of samples with label l, and

2. Generate atoms θ_k from H, i.e,

$$\theta_k \quad \sim H. \quad (3)$$

The work of the first author was supported by NSF under Award CCF-1018323 and by the ONR under Award N00014-09-1-1154.

We represent the obtained realization by

$$p(\theta) \quad = \quad \sum_{k=1}^{\infty} w_k \delta_{\theta_k}, \tag{4}$$

where $\sum_{k=1}^{\infty} w_k = 1$.

For this process we have the well-known Chinese Restaurant Process (CRP) metaphor. It goes as follows: a customer comes to an empty Chinese restaurant and is seated at table 1 where one type of dish is served. The second customer comes in and the customer is seated at the first table with probability $1/(\alpha+1)$ or at the second table (that serves another dish) with probability $\alpha/(\alpha+1)$. The nth customer comes in when there are L tables occupied. The probability that a customer joins table i, where $i \leq L \leq n-1$ is given by

$$p_i \quad = \quad \frac{n_i}{\alpha+n-1}, \tag{5}$$

where n_i is the number of customers already seated at table i, and the probability of the customer being seated at a new table is

$$p_{L+1} \quad = \quad \frac{\alpha}{\alpha+n-1}. \tag{6}$$

The mean and the variance of $G(A)$ are given by $\mathbb{E}\big(G(A)\big) = H(A)$ and $\mathrm{Var}\big(G(A)\big) = H(A)(1-H(A))/(\alpha+1)$, respectively. Clearly, the base distribution is the mean of the DP, and the value of the concentration parameter α affects the variance of the DP.

We pointed out that for classification, one uses Dirichlet mixture models. The implementations of classification are iterative schemes where in each iteration the considered data are classified into an existing class or a new one according to probabilities that are computed from the DP mixture model. In computing the relevant probabilities, one needs to find the predictive distributions of the data conditioned on the considered class and the data already classified in that class. If we denote all the data with $\mathcal{Y}$ and there are n data $y_i, i = 1, 2, \cdots, n$, the predictive distribution of y_i is a mixture distribution given by

$$\begin{aligned} p(y_i|\mathcal{Y}_{-i}, \mathcal{D}_{-i}) \quad &\propto \quad \frac{\alpha}{n-1+\alpha}\, p(y_i|\text{new class}) \\ &+ \quad \sum_{c_j} \frac{n_{c_j,-i}}{n-1+\alpha}\, p(y_i|\mathcal{Y}_{c_j,-i}), \end{aligned} \tag{7}$$

where $\mathcal{D}_{-i}$ are the current decisions of classification made about all the data except for y_i, $\mathcal{Y}_{-i}$ is the set of all data except y_i, $\mathcal{Y}_{c_j}$ is the set of data classified in c_j, and $n_{c_j,-i}$ is the number of data in class c_j.

3. MODELS WITH MULTIPLE DIRICHLET PROCESSES

We propose the construction of a process from two "elementary" DPs, $\mathrm{DP}(\alpha_1, H_1)$ and $\mathrm{DP}(\alpha_2, H_2)$, as follows:

1. For $i = 1, 2$, generate labels independently by

$$z_{i,n}|z_{i,1}, \cdots, z_{i,n-1} \sim \begin{cases} p_{i,l} = \frac{n_{i,l}}{\alpha_i+n-1}, & l = 1, 2 \cdots, L_i \\ p_{i,L_i+1} = \frac{\alpha_i}{\alpha_i+n}, & l = L_i+1, \end{cases} \tag{8}$$

   where $i = 1, 2$.

2. Generate atoms $\theta_{i,k}$ by

$$\theta_{i,k} \quad \sim \quad H_i, \quad i = 1, 2. \tag{9}$$

We represent the obtained realization by

$$p(\theta) \quad = \quad \sum_{k_1=1}^{\infty} \sum_{k_2=1}^{\infty} w_{1,k_1} w_{2,k_2} \delta_{\theta_{1,k_1}, \theta_{2,k_2}} \tag{10}$$

$$= \quad \sum_{k_1=1}^{\infty} w_{1,k_1} \delta_{\theta_{1,k_1}} \sum_{k_2=1}^{\infty} w_{2,k_2} \delta_{\theta_{2,k_2}}, \tag{11}$$

where $\sum_{k_1=1}^{\infty} \sum_{k_2=1}^{\infty} w_{1,k_1} w_{2,k_2} = 1$. We can show that this process is *not* a DP; however, its marginals are DPs.

For this process we have a modified CRP metaphor. A customer comes to an empty Chinese restaurant where the tables are ordered on a grid. The customer is seated at table 1 (located at the northwest corner of the restaurant, and denoted by $(1,1)$), where two types of dishes are served. The second customer comes in and the customer is seated at the same table $(1,1)$ as the first customer with probability

$$p_{1,1} \quad = \quad \frac{1}{(\alpha_1+1)(\alpha_2+1)}, \tag{12}$$

or at the second table in the first row, $(1,2)$, with probability

$$p_{1,2} \quad = \quad \frac{\alpha_2}{(\alpha_1+1)(\alpha_2+1)}, \tag{13}$$

or at the second table in the first column, $(2,1)$, with probability

$$p_{2,1} \quad = \quad \frac{\alpha_1}{(\alpha_1+1)(\alpha_2+1)}, \tag{14}$$

or at table $(2,2)$ with probability

$$p_{2,2} \quad = \quad \frac{\alpha_1\alpha_2}{(\alpha_1+1)(\alpha_2+1)}. \tag{15}$$

Suppose that by the time customer n comes in, the maximum row with an occupied table is L_1 and the maximum column is L_2. The customer may join any of the tables from row 1 to row L_1+1 and column 1 to column L_2+1. We define the probability that the customer will be seated at a table in row i by

$$p_{i,*} \quad = \quad \frac{n_{i,*}}{\alpha_1+n-1}, \quad i \leq L_1, \tag{16}$$

where $n_{i,*}$ is the total number of customers sitting in row i, or

$$n_{i,*} \quad = \quad \sum_{j=1}^{L_2} n_{i,j}, \tag{17}$$

and the probability that the table will be in row L_1+1 by

$$p_{L_1+1,*} \quad = \quad \frac{\alpha_1}{\alpha_1+n-1}. \tag{18}$$

Similarly, we have that a customer will sit at a table in column j with probability

$$p_{*,j} \quad = \quad \frac{n_{*,j}}{\alpha_2+n-1}, \quad j \leq L_2, \tag{19}$$

where $n_{*,j}$ is the total number of customers sitting in column j, i.e.,

$$n_{*,j} \quad = \quad \sum_{i=1}^{L_1} n_{i,j}, \tag{20}$$

and the probability that the table will be in column L_2+1 by

$$p_{*,L_2+1} \quad = \quad \frac{\alpha_2}{\alpha_2+n-1}. \tag{21}$$

Then the probability that the customer will join a table in row i and column j, where $i \leq L_1 + 1$, $j \leq L_2 + 1$, is given by

$$p_{ij} \quad = \quad p_{i,*}p_{*,j}. \tag{22}$$

More specifically, we can write

$$p_{i,j} \quad = \quad = \begin{cases} \frac{n_{i,*}n_{*,j}}{(\alpha_1+n-1)(\alpha_2+n-1)}, & i \leq L_1, j \leq L_2 \\ \frac{\alpha_1 n_{*,j}}{(\alpha_1+n-1)(\alpha_2+n-1)}, & i = L_1+1, j \leq L_2 \\ \frac{n_{i,*}\alpha_2}{(\alpha_1+n-1)(\alpha_2+n-1)}, & i \leq L_1+1, j = L_2 \\ \frac{\alpha_1\alpha_2}{(\alpha_1+n-1)(\alpha_2+n-1)}, & i = L_1+1, j = L_2+1. \end{cases} \tag{23}$$

Clearly, we note that the process defined here has a richer structure than a DP(α, H) that produces atoms according to

$$p(\theta) \quad = \quad \sum_{k=1}^{\infty} w_k \delta_{\theta_{1,k},\theta_{2,k}}. \tag{24}$$

In the CRP described by (10), two different dishes are served on each table, and the process allows the tables to have the same first dish but different second dish and vice versa, the same second dish but different first dish. By contrast, the process (24), which formally also has two dishes per table, does not have this flexibility. In other words, the process (24) only occupies the tables (i, i) on the diagonal of the restaurant, whereas the process (10) can occupy off-diagonal tables too. We note that with the parameters α_1 and α_2, we can tune how quickly the numbers of new rows and columns, respectively, grow with the arrival of new customers.

The process defined here can further be generalized by adding more dishes on every table. The introduction of mixture models with these processes that can be used for classification is straightforward. The classification can be implemented iteratively adapting the algorithm #3 described in [10].

4. AN EXAMPLE

We present the use of the new model for classification of multivariate data series.

Suppose we observe a set of multivariate data series Y_k, $k = 1 : n$ where $Y_k \in \mathbb{R}^{\ell \times t}$ is in class $c_{i,j}$ when,

$$Y_k = \Theta_{c_{i,*}} X_k + \Omega_{c_{*,j}} Z_k + U_k, \ k = 1 \ldots n,$$

where

- $\Theta_{c_{i,*}} \in \mathbb{R}^{\ell \times q}$ and $\Omega_{c_{*,j}} \in \mathbb{R}^{\ell \times q}$ are unknown matrices,
- $X_k \in \mathbb{R}^{q \times t}$ and $Z_k \in \mathbb{R}^{q \times t}$ are known,
- $U_k \in \mathbb{R}^{l \times t}$ is a random matrix of model errors where U_k is distributed according to a matrix normal distribution [12], i.e.,

  $$U_k \sim \text{MN}_{\ell,N}(U|0, \sigma^2 I).$$

  We consider that a matrix $M \in \mathbb{R}^{m \times l}$ is matrix normal distributed with mean μ and covariance Σ, noted as $M \sim \text{MN}_{m,l}(M|\mu, \Sigma)$, if the columns of M are independent and normally distributed with covariance Σ.

  The U_k are assumed independent and σ^2 is known.

In the last part of this section, we describe the steps in computing the predictive distribution $p(Y_k|\mathcal{Y}_{-k}, \mathcal{C}_{-k}, c_{i,j})$: the predictive distribution of Y_k, computed at Y_k, given the set of data records $\mathcal{Y}_{-k} = \{Y_i, \ i = 1 : n, \ i \neq k\}$, their corresponding classes $\mathcal{C}_{-k}$, and assuming that Y_k belongs to class $c_{i,j}$.

$$p(Y_k|\mathcal{Y}_{-k}, \mathcal{C}_{-k}, c_{i,j}) = \int F(Y_k; \phi_{c_{i,j}}) p(\phi_{c_{i,j}}|\mathcal{Y}_{-k}, \mathcal{C}_{-k}) d\phi_{c_{i,j}}, \tag{25}$$

where in our case $\phi_{c_{i,j}} = (\Theta_{c_{i,*}}, \Omega_{c_{*,j}})$ and $F(Y_i; \phi_{c_{i,j}})$ denotes the likelihood of Y_i. It is important to note that in computing the posterior distribution $p(\phi_{c_{i,j}}|\mathcal{Y}_{-k}, \mathcal{C}_{-k})$ we use the base measure of the DP process H as a prior.

More specifically, we assume H is a product of matrix normal distributions, i.e.,

$$(\Theta_{c_{i,*}}, \Omega_{c_{*,i}}) \sim \text{MN}_{\ell,q}(\Theta_{c_{i,*}}|0, \eta^2 I)\text{MN}_{\ell,q}(\Omega_{c_{*,i}}|0, \eta^2 I), \tag{26}$$

where the parameter of the prior η^2 is assumed known.

We denote by $Y_{(r,s)}$, $X_{(r,s)}$, $Z_{(r,s)}$ the $\ell \times t_{(r,s)}$ and $q \times t_{(r,s)}$ matrices obtained by concatenation of the Y_q, X_q and Z_q, $q \neq k$, classified in $c_{r,s}$, i.e., with parameters (Θ_r, Ω_s), where we have assumed without loss of generality that $c_{r,*} = r$, $r \in \{1 \ldots, L_1\}$ and $c_{*,s} = s$, $s \in \{1 \ldots, L_2\}$. We have,

$$Y_{(r,s)}|\Theta_r, \Omega_s \sim MN_{\ell,t_{(r,s)}}(Y|\Theta_r X_{(r,s)} + \Omega_s T_{(r,s)}, \sigma^2 I).$$

It is then possible to merge all the elements of $\mathcal{Y}_{-k}$ in an $\ell \times (n-1)t$ matrix as follows:

$$Y = (Y_{(1,1)}, \ldots, Y_{(1,L_2)}, Y_{(2,1)}, \ldots, Y_{(2,L_2)}, Y_{(3,1)} \ldots).$$

We can write in matrix form

$$Y_{-k} = \Gamma T_{-k} + U_{-k}, \tag{27}$$

where

$$\Gamma = (\Theta_1, \ldots, \Theta_{L_1}, \Omega_1, \ldots, \Omega_{L_2}), \ T_{-k} = \begin{pmatrix} T_1 \\ T_2 \end{pmatrix},$$

$$T_1 = \begin{pmatrix} X_{(1,1)} & \cdots & X_{(1,L_2)} & 0 & \cdots & & \\ 0 & \cdots & 0 & X_{(2,1)} & \cdots & X_{(2,L_2)} & \\ 0 & \cdots & & & \cdots & 0 & \cdots \\ \vdots & & & & & \vdots & \cdots \end{pmatrix},$$

$$T_2 = \begin{pmatrix} Z_{(1,1)} & 0 & \cdots & 0 & Z_{(2,1)} & 0 & \cdots \\ 0 & Z_{(1,2)} & \ddots & & \ddots & Z_{(2,2)} & \ddots \\ \vdots & \ddots & \ddots & \ddots & & \ddots & \ddots \\ 0 & \cdots & 0 & Z_{(1,L_2)} & 0 & \cdots & 0 \end{pmatrix}.$$

Note that if the class $c_{r,s}$ is empty, the corresponding $Y_{(r,s)}$ in Y and the columns of T_1 and T_2 containing $X_{(r,s)}$ and $Z_{(r,s)}$ are removed. If the number of non-empty classes is $M \leq L_1 L_2$, the number of block matrices in T_1 is $L_1 \times M$ and in T_2 is $L_2 \times M$. The size of T_1 is $qL_1 \times (n-1)t$ and the size of T_2 is $qL_2 \times (n-1)t$.

With this notation,

$$\Gamma \sim \text{MN}_{\ell,q(L_1+L_2)}(0, \eta^2 I, I), \tag{28}$$

$$U \sim \text{MN}_{\ell,(n-1)t}(0, \sigma^2 I, I). \tag{29}$$

Similarly to (27), we can write

$$Y_k = \Gamma T_k + U_k \tag{30}$$

where T_k is a $q(L_1 + L_2) \times t$ matrix constructed in the same way as T_{-k} with X_k and Z_k in the appropriate place.

The application of classical results of Bayesian linear regression, see for example [13], to model (27), (30) with assumptions (28,29) shows that the predictive distribution of $Y_k|Y_{-k}$ is multivariate normal. More specifically,

$$Y_k|\mathcal{Y}_{-k}, \mathcal{C}_{-k}, c_{i,j} \sim \text{MN}_{Y_k|\ell,t}(\mu_p, I, \Sigma_p),$$

where

$$\mu_p = \mu_\Gamma T_k, \quad \Sigma_p = \sigma^2 I + T_k^\top \Sigma_\Gamma T_k.$$

The matrices μ_Γ and Σ_Γ are the mean and covariance of the posterior distribution $\Gamma|Y \sim \text{MN}_{\ell,q(L_1+L_2)}(\Gamma|\mu_\Gamma, \Sigma_\Gamma)$ with

$$\Sigma_\Gamma = (\sigma^{-2} T_{-k} T_{-k}^\top + \eta^{-2} I)^{-1}, \quad \mu_\Gamma = \sigma^{-2} \Sigma_\Gamma Y_{-k} T_{-k}^\top.$$

The computation of Σ_Γ and μ_Γ require computation of $T_{-k}T_{-k}^\top$ and $Y_{-k}T_{-k}$, which involve the large matrices T_{-k} and Y_{-k}. These matrix products can be realized efficiently using the block structure of T_{-k}. Compact expressions are not reported here for lack of space.

The classification algorithm will also require the computation of the predictive distribution (25) when the classes $c_{i,*}$ or/and $c_{*,j}$ are empty.

5. SIMULATION RESULTS

The hyperparameters of the prior and the variance of the noise were set to $\eta^2 = 1$ and $\sigma^2 = 1.5$ with $\ell = 2$, $q = 3$, $t = 5$, and $n = 9$. The number of iterations of the Gibbs sampler was fixed to 500 and the burn-in period to 50.

Experiment #1

In the first experiment, the data were generated according to three classes which are detailed in the following table:

PARAMETERS OF THE 3 CLASSES – EXPERIMENT # 1		
Y_i $i = 1:3$	$\Theta_1 \sim \mathcal{MN}_{2,3}(\Theta\|0, I)$	$\Omega_1 \sim \mathcal{MN}_{2,3}(\Theta\|0, I)$
Y_i $i = 4:6$	Θ_1	$\Omega_2 \sim \mathcal{MN}_{2,3}(\Theta\|0, I)$
Y_i $i = 7:9$	$\Theta_2 \sim \mathcal{MN}_{2,3}(\Theta\|0, I)$	$\Omega_3 \sim \mathcal{MN}_{2,3}(\Theta\|0, I)$

The two concentration parameters were $\alpha_1 = \alpha_2 = 0.1$. The results obtained from a single realization are presented in Fig. 1. The first three plots are the histograms of the number of classes w.r.t. Θ, Ω and (Θ, Ω). The histograms were obtained from the samples of the Markov chain. The third plot shows the "confusion matrix" estimated from the samples of $c_1, \ldots, c_n|Y$. The pixel (i, j) in the plot represents the probability that Y_i and Y_j are in the same class.

These plots clearly show that the number of classes is correctly estimated, and the classification of Y_i is correct. The left plot of Fig. 2 gives the results obtained under the same conditions with a standard DP. Both classifiers behave similarly.

The right plot of Fig. 2 gives the result obtained with a PDP mixture model when $\alpha_1 = 0.1$ and $\alpha_2 \approx 0$. In this case, as expected, the classification is performed only w.r.t. Θ, leading to 2 classes: the first one contains $\{Y_i,\ i = 1:6\}$ and the second one $\{Y_i,\ i = 7:9\}$. This possibility clearly shows the flexibility of our model.

Experiment #2

In the second experiment, the parameters of the three classes were

PARAMETERS OF THE 3 CLASSES – EXPERIMENT # 1		
Y_i $i = 1:3$	$\Theta_1 \sim \mathcal{MN}_{2,3}(\Theta\|0, I)$	$\Omega_1 \sim \mathcal{MN}_{2,3}(\Theta\|0, I)$
Y_i $i = 4:6$	Θ_1	$\Omega_2 = 2\Omega_1$
Y_i $i = 7:9$	$\Theta_2 \sim \mathcal{MN}_{2,3}(\Theta\|0, I)$	$\Omega_3 \sim \mathcal{MN}_{2,3}(\Theta\|0, I)$

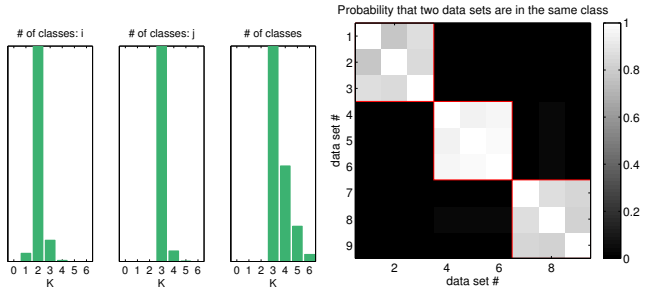

Fig. 1: EXPERIMENT #1. Left: Histogram of the number of classes. Right: Estimated posterior probability of Y_i and Y_j being in the same class. The red boxes indicate the true classes. $\alpha_1 = \alpha_2 = 0.1$.

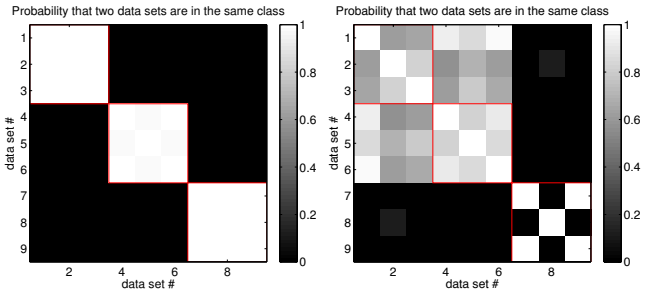

Fig. 2: EXPERIMENT #1. Left: Classification with a standard DP, $\alpha = 0.1$. Right: Classification with a PDP with $\alpha_1 = 0.1$, $\alpha_2 \approx 0$ in order to classify w.r.t. Θ.

The left plot of Fig. 3 gives the result obtained with a DP and $\alpha = 0.1$. Similar results are obtained with a PDP using parameters $\alpha_1 = \alpha_2 = 0.1$. The classifiers show difficulties *with these parameters*. The right plot of Fig. 3 gives the result obtained with a PDP and $\alpha_1 = 0.1$, $\alpha_2 = 1$. This set of parameters allows a correct separation of the 3 classes.

6. CONCLUSION

In this paper, we proposed products of DPs as a basis for building mixture models for classification. With these models we allow for sharing of mixture components of classes of data. We showed how this modeling can be used for classification of multivariate time series. The obtained results were used in computer simulations.

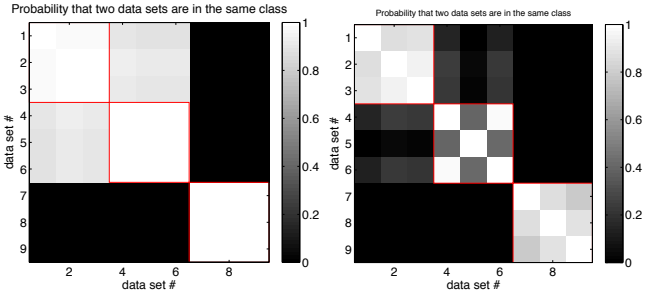

Fig. 3: EXPERIMENT #2. Left: Classification with a standard DP with $\alpha = 0.1$ or a PDP with $\alpha_1 = \alpha_2 = 0.1$. Right: Classification with a PDP with $\alpha_1 = 0.1$, $\alpha_2 = 1$.

7. REFERENCES

[1] D. Blei, L. Carin, and D. Dunson, "Probabilistic topic models," *Signal Processing Magazine*, vol. 27, no. 6, pp. 55–65, 2010.

[2] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "A sticky HDP-HMM with application to speaker diarization," *Annals of Applied Statistics*, vol. 5, no. 2A, pp. 1020–1056, 2011.

[3] T. L. Griffiths and Z. Ghahramani, "The Indian buffet process: An introduction and review," *Journal of Machine Learning Research*, pp. 1185–1224, 2011.

[4] T. S. Ferguson, "A Bayesian analysis of some nonparametric models," *Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[5] Y. W. Teh, "Dirichlet processes," *Encyclopedia of Machine Learning*, Springer, 2010.

[6] C. Antoniak, "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *Annals of Applied Statistics*, vol. 2, no. 6, pp. 1152–1174, 1974.

[7] M. D. Escobar and M. West, "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, vol. 430, no. 6, pp. 577–588, 1995.

[8] T. Ferguson, "Bayesian density estimation by mixtures of normal distributions," in *Recent Advances in Statistics*, M. Rizvi, J. Rustagi, and D. Siegmund, Eds., pp. 287–302. Academic Press, 1983.

[9] N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, Eds., *Bayesian Nonparametrics*, Cambridge University Press, 2010.

[10] R. Neal, "Markov chain sampling methods for Dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 249–265, 2000.

[11] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[12] A. K. Gupta and D. K. Nagar, *Matrix variates distributions*, Chapman and Hall/CRC, 2000.

[13] G. E. P. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis*, Wiley, 2011.