

# Dirichlet process mixture models for finding shared structure between two related data sets

Gayle Leen  
University of Paisley  
School of Computing  
PA1 2BE  
Scotland

Colin Fyfe  
University of Paisley  
School of Computing  
PA1 2BE  
Scotland

*Abstract:* A nonparametric Bayesian approach is used for the problem of learning from two related data sets. We model the shared structure between two data sets using a Dirichlet process mixture model of probabilistic canonical correlation analysers, which allows the flexibility of the mappings from shared feature to data spaces to be automatically determined from the data.

*Keywords:* Nonparametric Bayesian methods, Dirichlet processes, Probabilistic Canonical Correlation Analysis, Mixture models

## 1 Introduction

In general, research in the machine learning field has focused on analysing data that is the output of a single sensor (a single data source) rather than analysing data from the output of several sensors. However, it seems advantageous to learn from multiple data sources because there is more information about the underlying data generating process than if we had just considered a single source. The relevance of this research area is founded in the human's brain's ability to integrate five different sensory input streams into a coherent representation of its environment. Additionally, due to the increased availability of electronic recording devices and advances in data analysis techniques, there exists many scenarios in which it becomes necessary to model multiple data sources.

### 1.1 Learning from two data sources

In this paper, we present a model for finding a joint probabilistic representation of two data sources, which builds on work in [1]. In general, existing methods for finding shared structure are discriminative methods, which find a set of features for each set that optimise a similarity measure between the features e.g. [2],[3], [4]. Using these methods can be problematic; a probability density is not defined over the two sets of data variables, and therefore we cannot evaluate quantities such as the predictive density over one data set given the other. Additionally, these methods do not model the underlying data generating process. Though this may be efficient in that the modelling power is focused on optimising the quantity of interest - the similarity of the extracted features - it is difficult to incorporate prior knowledge about the feature space. With

this lack of knowledge about the problem, care has to be taken in designing appropriate nonlinear mappings for finding nonlinearly related pairs of features using discriminative techniques. An inflexible mapping may not recover the true underlying shared structure between the data sets, and an overly flexible mapping may find spurious correlations between the data sets.

This problem of inferring the appropriate complexity of the model can be addressed using nonparametric Bayesian methods. The complexity of the model is allowed to grow with the number of data points such that the necessary complexity is inferred from the data. This involves placing a prior over a family of probability distributions over the data generating process to allow a flexible prior on the underlying data distribution. One such prior from the nonparametric statistics field is the Dirichlet process (DP) [5], which is a distribution over distributions. In this paper, we assume that each data set lies close to a nonlinear manifold in data space, each indexed by a shared set of latent coordinates, which reflects the shared structure underlying the data sets. We extend the probabilistic formulation of canonical correlation analysis (PCCA) [6] to a mixture of PCCA in the spirit of the mixture of probabilistic principal component analyzers [7] to find a low dimensional representation of two related data sources. The resulting model approximates the pair of nonlinear manifolds by pairs of local linear submodels. We use the DP as a nonparametric prior for the parameters of the mixture model, allowing the number of mixture components to grow with the number of data points, such that the flexibility of the manifolds is inferred from the data automatically. We call this model a Dirichlet process mixture model of probabilistic canonical correlation analysers.

## 1.2 Canonical correlation analysis

Canonical correlation analysis (CCA) [2] is concerned with finding linear relationships between the two sets of variables. Given two sets of zero mean data variables  $\mathbf{y}_1 \in \mathfrak{R}^{m_1}$  and  $\mathbf{y}_2 \in \mathfrak{R}^{m_2}$ , where  $m_1$  and  $m_2$  are the dimensions of  $\mathbf{y}_1$  and  $\mathbf{y}_2$  respectively, CCA finds linear projections of each variable  $\mathbf{x}_1 = \mathbf{U}_1^\top \mathbf{y}_1$  and  $\mathbf{x}_2 = \mathbf{U}_2^\top \mathbf{y}_2$ , termed the canonical variates, such that the correlation between  $\mathbf{x}_1$  and  $\mathbf{x}_2$  is maximised, and  $\mathbf{U}_1 \in \mathfrak{R}^{m_1 \times q}$  and  $\mathbf{U}_2 \in \mathfrak{R}^{m_2 \times q}$ , where  $q \leq \min(m_1, m_2)$ , are matrices whose columns  $\mathbf{U}_{1,i}, \mathbf{U}_{2,i}, i = 1, \dots, q$  form the  $q$  pairs of canonical vectors. We can find  $\mathbf{U}_1$  and  $\mathbf{U}_2$  as the eigenvectors of the generalised eigenvalue problem:

$$\begin{bmatrix} 0 & \tilde{\Sigma}_{12} \\ \tilde{\Sigma}_{21} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{bmatrix} = \begin{bmatrix} \tilde{\Sigma}_{11} & 0 \\ 0 & \tilde{\Sigma}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{bmatrix} \rho$$

where  $\rho$  is the diagonal matrix of canonical correlations, and

$$\tilde{\Sigma} = E \left[ \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix}^\top \right] = \begin{bmatrix} \tilde{\Sigma}_{11} & \tilde{\Sigma}_{12} \\ \tilde{\Sigma}_{21} & \tilde{\Sigma}_{22} \end{bmatrix}$$

Another property of CCA is that the projections onto canonical directions corresponding to a different canonical correlation are uncorrelated such that  $\mathbf{U}_1^\top \tilde{\Sigma}_{11} \mathbf{U}_1 = \mathbf{I}_{m_1}$  and  $\mathbf{U}_2^\top \tilde{\Sigma}_{22} \mathbf{U}_2 = \mathbf{I}_{m_2}$ .

## 1.3 Probabilistic canonical correlation analysis

Canonical correlation analysis (CCA) was formulated as a latent variable model in [6]. It is found that the posterior distributions of the latent variables lie in the same linear subspaces as those defined by standard CCA. The likelihood for a pair of data points is given by:

$$p(\mathbf{y} | \mathbf{x}) = \mathcal{N}(\mathbf{y} | \mathbf{W}\mathbf{x} + \boldsymbol{\mu}, \boldsymbol{\Psi}) \quad (1)$$

$\mathbf{y}$  is defined as the concatenation of two sets of data variables i.e.  $\mathbf{y} = [\mathbf{y}_1^\top \mathbf{y}_2^\top]^\top$ , where  $\mathbf{y}_1 \in \mathfrak{R}^{m_1}, \mathbf{y}_2 \in \mathfrak{R}^{m_2}$  with  $m_1$  and  $m_2$  being the dimensions of the two data variable sets,  $\mathbf{W} = [\mathbf{W}_1^\top \mathbf{W}_2^\top]^\top$  with  $\mathbf{W}_1 \in \mathfrak{R}^{m_1 \times q}, \mathbf{W}_2 \in \mathfrak{R}^{m_2 \times q}$ , and  $\boldsymbol{\mu}$  is the bias parameter.  $\mathbf{x}_n \in \mathfrak{R}^q$  (where  $q$  is the latent space dimensionality) is the shared latent variable for the  $n$ th pair of data variables  $\mathbf{y}_n$ , and distributed according to  $\mathbf{x} \sim \mathcal{N}(\mathbf{x} | 0, \mathbf{I}_q)$ . The noise covariance matrix is constrained to be of block diagonal form:

$$\Sigma_n = \begin{pmatrix} \boldsymbol{\Psi}_1 & 0 \\ 0 & \boldsymbol{\Psi}_2 \end{pmatrix} \quad (2)$$

where  $\boldsymbol{\Psi}_1 \in \mathfrak{R}^{m_1 \times m_1} \succeq 0, \boldsymbol{\Psi}_2 \in \mathfrak{R}^{m_2 \times m_2} \succeq 0$ . The maximum likelihood solutions for the parameters are given by:

$$\hat{\mu}_1 = \tilde{\mu}_1 \quad (3)$$

$$\hat{\mu}_2 = \tilde{\mu}_2 \quad (4)$$

$$\hat{\mathbf{W}}_1 = \tilde{\Sigma}_{11}^{-1} \mathbf{U}_{1q} \mathbf{P}_q \mathbf{R} \quad (5)$$

$$\hat{\mathbf{W}}_2 = \tilde{\Sigma}_{22}^{-1} \mathbf{U}_{2q} \mathbf{P}_q \mathbf{R} \quad (6)$$

$$\hat{\boldsymbol{\Psi}}_1 = \tilde{\Sigma}_{11} - \hat{\mathbf{W}}_1 \hat{\mathbf{W}}_1^\top \quad (7)$$

$$\hat{\boldsymbol{\Psi}}_2 = \tilde{\Sigma}_{22} - \hat{\mathbf{W}}_2 \hat{\mathbf{W}}_2^\top \quad (8)$$

where  $\mathbf{U}_{1q} \in \mathfrak{R}^{m_1 \times q}$  and  $\mathbf{U}_{2q} \in \mathfrak{R}^{m_2 \times q}$  are matrices whose columns consist of the first  $q$  canonical directions for  $\mathbf{y}_1$  and  $\mathbf{y}_2$  respectively,  $\mathbf{P}_q$  is the diagonal matrix of the  $q$  largest canonical correlations,  $\mathbf{R} \in \mathfrak{R}^{q \times q}$  is a rotation matrix.

## 1.4 Dirichlet processes

The Dirichlet process (DP) is a nonparametric distribution on distributions, or equivalently, a measure on measures [5]. We can view a DP as an infinite dimensional Dirichlet distribution. A DP is parameterised by a scaling parameter  $\alpha_0 > 0$ , and a base measure  $G_0$ . Suppose that  $G_0$  is a distribution over a measurable space  $\Theta$ . This acts as the base measure for the DP. A Dirichlet process is defined to be the distribution of a random probability measure  $G$  over  $\Theta$  i.e.

$$G \sim \text{DP}(G | G_0, \alpha_0) \quad (9)$$

such that for any  $K$  finite partitions of  $\Theta$ ,  $\{A_1, \dots, A_K\}$ ,  $\{G(A_1), \dots, G(A_K)\}$  follows a finite dimensional Dirichlet distribution with parameters  $\{\alpha_0 G_0(A_1), \dots, \alpha_0 G_0(A_K)\}$ :

$$\{G(A_1), \dots, G(A_K)\} \sim \text{Dir}(\alpha_0 G_0(A_1), \dots, \alpha_0 G_0(A_K)) \quad (10)$$

where  $\alpha_0 > 0$  determines the concentration of  $\{G(A_1), \dots, G(A_K)\}$  around  $\{G_0(A_1), \dots, G_0(A_K)\}$ . As  $\alpha_0 \rightarrow \infty, G \rightarrow G_0$ . One perspective on the Dirichlet process is provided by the Pólya urn scheme [8], which demonstrates the clustering property of draws from  $G$ , in that a set of samples  $\{\Theta_1, \dots, \Theta_N\}$  drawn from  $G$  are not necessarily distinct. This means that the data is divided into  $K$  partitions, or clusters, where each partition has the same parameter setting  $\theta^i$ . The more often  $\theta^i$  is drawn, the more likely it is to be drawn in the future.  $\alpha_0$  controls the tendency to form clusters; if  $\alpha_0$  is very small, it is likely that there will be few clusters, and if  $\alpha_0$  is large, there will be many small clusters.

## 2 A Mixture of Canonical Correlation Analysers

A mixture model models the density for a data point  $\mathbf{y}_n$  as a weighted average of  $K$  latent variable model densities,

where  $K$  is the number of mixture components. The probability for  $\mathbf{y}_n$  is given by:

$$p(\mathbf{y}_n | \theta) = \sum_{k=1}^K p(\mathbf{y}_n | \theta_k, c_n = k) p(c_n = k | \boldsymbol{\pi}) \quad (11)$$

where  $c \in \{1, \dots, K\}$  is a discrete variable which indicates which latent variable model has been chosen,  $\boldsymbol{\pi} = [\pi_1, \dots, \pi_K]^\top$  is a vector of mixing proportions (such that  $\sum_{k=1}^K \pi_k = 1$ ).  $p(\mathbf{c} | \boldsymbol{\pi})$  is a multinomial distribution over  $\mathbf{c}$ , where  $\mathbf{c} = \{c_1, \dots, c_N\}$  is the set of indicators for all  $N$  data points, such that  $p(c_n = k | \boldsymbol{\pi}) = \pi_k$ . To simplify notation, we will write  $c_n = k$  as  $k$  from now on.  $p(\mathbf{y}_n | \theta_k, k)$  is the probability of  $\mathbf{y}_n$  under the  $k$ th latent variable model, with the corresponding set of parameters  $\theta_k$ , and  $\theta = \{\theta_1, \dots, \theta_K\}$  is the complete set of parameters. To create a mixture of probabilistic Canonical Correlation Analysers, the  $k$ th latent variable model density has the form:

$$\begin{aligned} p(\mathbf{y}_n | \theta_k, k) &= \int p(\mathbf{y}_n | \mathbf{x}_n, \theta_k, k) p(\mathbf{x}_n | k) d\mathbf{x}_n \\ &= \mathcal{N}(\mathbf{y}_n | \mu_k, \mathbf{W}_k \mathbf{W}_k^\top + \boldsymbol{\Psi}_k) \end{aligned} \quad (12)$$

where

$$p(\mathbf{y}_n | \mathbf{x}_n, \theta_k, k) = \mathcal{N}(\mathbf{y}_n | \mathbf{W}_k \mathbf{x}_n + \mu_k, \boldsymbol{\Psi}_k) \quad (13)$$

$$p(\mathbf{x}_n | k) = \mathcal{N}(\mathbf{x}_n | \mathbf{0}, \mathbf{I}_q) \quad (14)$$

with  $\mathbf{y}_n$  defined as the concatenation of two sets of data variables i.e.  $\mathbf{y}_n = [\mathbf{y}_{1,n}^\top \mathbf{y}_{2,n}^\top]^\top$ , where  $\mathbf{y}_{1,n} \in \mathbb{R}^{m_1}$ ,  $\mathbf{y}_{2,n} \in \mathbb{R}^{m_2}$  with  $m_1$  and  $m_2$  being the dimensions of the two data variable sets,  $\mathbf{W}_k = [\mathbf{W}_{1,k}^\top \mathbf{W}_{2,k}^\top]^\top$  with  $\mathbf{W}_{1,k} \in \mathbb{R}^{m_1 \times q}$ ,  $\mathbf{W}_{2,k} \in \mathbb{R}^{m_2 \times q}$ ,  $\mu_k$  is the bias parameter and  $\mathbf{x}_n \in \mathbb{R}^q$  is the corresponding shared latent variable. The noise covariance matrix is constrained to be of block diagonal form:

$$\boldsymbol{\Psi}_k = \begin{pmatrix} \boldsymbol{\Psi}_{1,k} & 0 \\ 0 & \boldsymbol{\Psi}_{2,k} \end{pmatrix} \quad (15)$$

where  $\boldsymbol{\Psi}_{1,k} \in \mathbb{R}^{m_1 \times m_1}$ ,  $\boldsymbol{\Psi}_{2,k} \in \mathbb{R}^{m_2 \times m_2}$ . We have assumed that the prior on the latent variable is the same for all  $K$  mixture components and that each Gaussian cluster has the same intrinsic dimensionality  $q$ , so from now on we will omit the indicator variable when denoting the latent priors, and rewrite  $p(\mathbf{x}_n | k)$  as  $p(\mathbf{x}_n)$ . The generative model for the mixtures of probabilistic canonical correlation analyzers is shown in Figure 1. A pair of data points is generated by first choosing a submodel  $k$  according to  $p(c_n = k | \boldsymbol{\pi})$ , and then drawing from the  $k$ th PCCA model  $p(\mathbf{y}_n | \theta_k, c_n = k)$ . The log likelihood function for the model is given by:

$$\begin{aligned} \mathcal{L} &= \sum_{n=1}^N \ln p(\mathbf{y}_n | \theta) \\ &= \sum_{n=1}^N \ln \sum_{k=1}^K p(k | \boldsymbol{\pi}) \int p(\mathbf{y}_n | \mathbf{x}_n, \theta_k, k) p(\mathbf{x}_n) d\mathbf{x}_n \end{aligned} \quad (16)$$

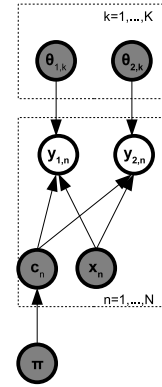


Figure 1: The generative model for the mixture of PCCA. A submodel  $k$  (indicated by  $c_n$ ) is chosen by drawing from  $p(c_n | \boldsymbol{\pi})$ , and  $\mathbf{x}_n$ , the shared latent variable is drawn from  $p(\mathbf{x})$ . Given  $c_n$ ,  $\mathbf{x}_n$  and the corresponding set of parameters  $\theta_{1,k} = \{\mathbf{W}_{1,k}, \mu_{1,k}, \boldsymbol{\Psi}_{1,k}\}$  and  $\theta_{2,k} = \{\mathbf{W}_{2,k}, \mu_{2,k}, \boldsymbol{\Psi}_{2,k}\}$ , the  $n$ th pair of data variables  $\mathbf{y}_{1,n}$  and  $\mathbf{y}_{2,n}$  are drawn from  $p(\mathbf{y}_{1,n} | \mathbf{x}_n, \theta_{1,k})$  and  $p(\mathbf{y}_{2,n} | \mathbf{x}_n, \theta_{2,k})$  respectively.

The Expectation-Maximization (EM) algorithm [9] can be used to find maximum likelihood estimates of the model parameters.

## 2.1 Generalising to an infinite mixture model

The previous section uses a maximum likelihood (ML) approach to finding the parameters of the model, in which the model parameters are assigned specific values which correspond to a (local) maximum of the likelihood function. One problem with the maximum likelihood method is that the method does not take model complexity into account, and the data is more likely under more complex model structures, which leads to overfitting. The likelihood increases with the number of components in the model, such that the likelihood is maximised for the extreme case where each data point is attributed to a separate mixture component.

An elegant solution to the model selection problem is the nonparametric Bayesian approach; by integrating out those parameters whose cardinality scales with model complexity, more complex models are penalised since they can *a priori* model a greater range of data sets. There are several Bayesian approaches to mixture modelling in the literature which approximate the integrals required for Bayesian inference, using sampling techniques [10, 11], and variational approximations [12, 13].

In these models, the number of components is found automatically. One approach is to set a maximum number of potential components, and then when the model is trained to some data, unwanted components are suppressed, such as in [13]. Similarly, in [12], variational approximations to a full Bayesian integration over the model parameters are derived for a Bayesian mixture of factor an-

alyzers. The model is initialized with a single component, and the number of components that fit the training data is found by adding new components through a stochastic procedure, and removing zero responsibility components when necessary.

Another way to approach the model selection problem regarding the number of components is to use Bayesian models with an infinite number of parameters such as the infinite mixture of Gaussians in [11]. This allows the model to be of the necessary complexity through considering a continuum of models and averaging with respect to all of these simultaneously, rather than controlling the complexity through limiting the number of parameters. Modelling data as coming from an infinite mixture has been seen to work well in the infinite mixture of Gaussians when there are only a small finite number of components in the actual mixture. The infinite mixture of Gaussians is similar to existing models in nonparametric statistics known as Dirichlet process mixture models [5, 14, 15] but derives the model as a limiting case of a finite mixture model rather than from the Dirichlet process itself such as in [16].

In the following section, we derive the Dirichlet process mixture model as the limiting case of the finite mixture model. Suppose that we place a symmetric Dirichlet prior on the mixing proportions of the  $K$  component mixture model  $\boldsymbol{\pi} = \{\pi_1, \dots, \pi_K\}$ , which is conjugate to the multinomial  $p(\mathbf{c} | \boldsymbol{\pi})$ , the distribution over the indicator variables  $\mathbf{c} = \{c_1, \dots, c_N\}$ :

$$p(\boldsymbol{\pi} | \alpha_0) = \text{Dir}(\boldsymbol{\pi} | \frac{\alpha_0}{K}, \dots, \frac{\alpha_0}{K}) = C(\alpha_0) \prod_{k=1}^K \pi_k^{\alpha_0/K - 1}$$

where  $\alpha_0 > 0$  is a positive scaling parameter,  $C(\alpha_0) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0/K)^K}$  is a normalisation constant, where  $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$  denotes the Gamma function, and  $\mathcal{E}(\pi_k) = 1/K$ . Integrating out the mixing proportions we get:

$$\begin{aligned} p(c_1, \dots, c_N | \alpha_0) &= \int p(\mathbf{c} | \boldsymbol{\pi}) p(\boldsymbol{\pi} | \alpha_0) d\boldsymbol{\pi} \\ &= \frac{\Gamma(\alpha_0)}{\Gamma(N + \alpha_0)} \prod_{k=1}^K \frac{\Gamma(N_k + \alpha_0/K)}{\Gamma(\alpha_0/K)} \end{aligned} \quad (17)$$

It is difficult to directly sample  $\mathbf{c}$  from this distribution; instead, the indicators are Gibbs sampled to capture their dependencies. The conditional prior over the indicator variable for the  $n$ th data point given all the other indicator variables is given by:

$$p(c_n = k | \mathbf{c}_{-n}, \alpha_0) = \frac{N_{-n,k} + \alpha_0/K}{N - 1 + \alpha_0} \quad (18)$$

where  $\mathbf{c}_{-n}$  denotes the set of indicators not including  $c_n$ , and  $N_{-n,k}$  is the number of data points in the  $k$ th cluster,

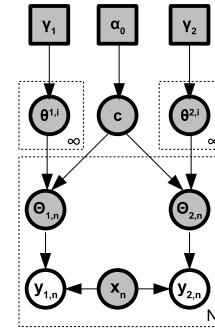


Figure 2: Graphical model for DP mixture model of PCCA

not including the  $n$ th data point. If we allow  $K \rightarrow \infty$ , i.e. we allow an infinite number of mixture components, the conditional prior on  $c_n$  becomes:

$$p(c_n = k | \mathbf{c}_{-n}, \alpha_0) = \frac{N_{-n,k}}{N - 1 + \alpha_0} \quad (19)$$

$$p(c_n \neq c_{n'} \forall n' \neq n | \mathbf{c}_{-n}, \alpha_0) = \frac{\alpha_0}{N - 1 + \alpha_0} \quad (20)$$

where the last equation is the probability that the data point is assigned to a new cluster. The parameters  $\{\Theta_1, \dots, \Theta_N\}$  for the data points are generated according to:

$$p(\Theta_1, \dots, \Theta_N | \theta, \alpha_0) = \sum_{\mathbf{c}} \left( \prod_{n=1}^N p(\Theta_n | c_n, \theta) \right) p(\mathbf{c} | \alpha_0)$$

This involves a summation over  $\mathbf{c}$  i.e. over all possible assignments of data points to the components, but it is easier to evaluate in terms of the Gibbs sampling scheme as in (17), and if  $c_n$  takes on an existing value, then the data point  $n$  inherits the parameter set  $\theta^{c_n}$ :  $\Theta_n = \theta^{c_n}$ . If  $c_n$  takes on a new value (starts a new cluster) then the parameter set is generated from the prior  $p(\theta | h)$ , where  $h$  is the set of hyperparameters. This is equivalent to the Pólya urn sampling scheme. This model is a Dirichlet process mixture model.

### 3 An infinite mixture of probabilistic CCA

In this section, we describe the Dirichlet process mixture model of probabilistic CCA, which uses a Dirichlet process prior on the parameters for each data point, as detailed in the previous sections.

#### 3.1 Overview of the model

This is equivalent to placing a DP prior on the indicators  $\mathbf{c} = \{c_1, \dots, c_N\}$  (which show the latent submodel with which the  $N$  pairs of data points are associated), and integrating over the mixing proportions  $\boldsymbol{\pi}$ . Priors are placed on the component parameters  $\theta_k$ . The graphical model

is shown in Figure 2. The probability of the data set  $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_N]^\top$  is given by:

$$p(\mathbf{Y} | \alpha_0, \gamma) = \prod_{n=1}^N \sum_{\mathbf{c}} \int p(\mathbf{y}_n | c_n, \theta) p(\mathbf{c} | \alpha_0) p(\theta | \gamma) d\theta \quad (21)$$

where  $p(\theta | \gamma)$  is the distribution over the parameter space  $\theta$  (equivalent to  $G_0$ ), with hyperparameters  $\gamma$ . This is chosen to be a conjugate prior to the probabilistic CCA likelihood.  $p(\mathbf{c} | \alpha_0)$  is the distribution over the indicator variables, where the conditional priors are given in (19) and (20), the Pólya urn scheme.  $p(\mathbf{y}_n | c_n, \theta)$  is the likelihood for a data point under the  $c_n$ th latent submodel in the probabilistic CCA model. When  $c_n = k$ , this is written as:  $p(\mathbf{y}_n | c_n = k, \theta^k) = \int p(\mathbf{y}_n | \mathbf{x}_n, \theta^k, c_n = k) p(\mathbf{x}_n) d\mathbf{x}_n$ . We can write the probability of the data set in terms on the  $K$  represented clusters:

$$p(\mathbf{Y} | \alpha_0, \gamma) = \sum_{\mathbf{c}} \prod_{k=1}^K \left( \int p(\mathbf{Y}^k | \theta^k, \mathbf{c}) p(\theta^k | \gamma) d\theta^k \right) p(\mathbf{c} | \alpha_0) \quad (22)$$

where  $p(\mathbf{Y}^k | \theta^k, \mathbf{c})$  is the probability of all the data pairs assigned to the  $k$ th cluster, given the assignments  $\mathbf{c}$  of all the data, parameterised by  $\theta^k$ . Additionally, we define separate parameters and hyperparameters for the two data sets  $\mathbf{Y}_1$  and  $\mathbf{Y}_2$  such that we can write:

$$p(\mathbf{Y} | \alpha_0, \gamma) = p(\mathbf{Y}_1 | \alpha_0, \gamma_1) p(\mathbf{Y}_2 | \alpha_0, \gamma_2) \quad (23)$$

where for  $i = 1, 2$

$$p(\mathbf{Y}_i | \alpha_0, \gamma_i) = \sum_{\mathbf{c}} \prod_{k=1}^K \left( \int p(\mathbf{Y}_i^k | \theta^{i,k}, \mathbf{c}) p(\theta^{i,k} | \gamma_i) d\theta^{i,k} \right) p(\mathbf{c} | \alpha_0)$$

where  $\mathbf{Y}_i^k$  is the  $k$ th cluster of the  $i$ th data set,  $\theta^{i,k}$  is the set of parameters for the  $k$ th latent submodel for the  $i$ th data set, governed by the set of hyperparameters  $\gamma_i$ . With this formulation, it is easy to see how to compute the posterior distributions over the indicators  $\mathbf{c}$ , the parameters  $\theta = \{\theta^1, \dots, \theta^K\}$ , and the hyperparameters  $\gamma$  and  $\alpha_0$ .

### 3.1.1 Posterior over the parameters

The posterior distributions over the  $k$ th set of parameters are given by:

$$p(\theta^{1,k} | \mathbf{Y}^{1,k}, \mathbf{c}, \gamma_1) \propto p(\mathbf{Y}^{1,k} | \theta^{1,k}, \mathbf{c}) p(\theta^{1,k} | \gamma_1) \quad (24)$$

$$p(\theta^{2,k} | \mathbf{Y}^{2,k}, \mathbf{c}, \gamma_2) \propto p(\mathbf{Y}^{2,k} | \theta^{2,k}, \mathbf{c}) p(\theta^{2,k} | \gamma_2) \quad (25)$$

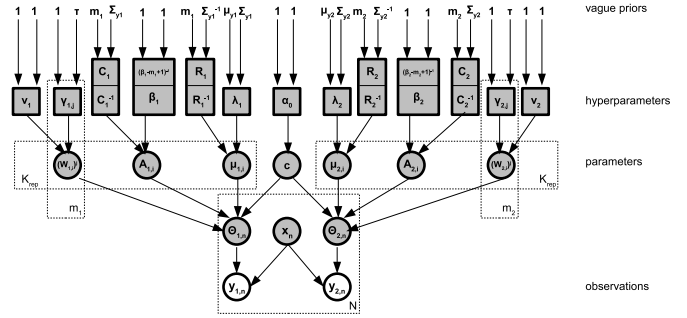


Figure 3: The complete graphical model for the Dirichlet process mixture model of probabilistic CCA.

### 3.1.2 Posterior over the hyperparameters

The posterior distributions over the hyperparameters given the  $K$  sets of parameters are:

$$p(\gamma_1 | \theta^{1,1}, \dots, \theta^{1,K}) \propto \prod_{i=1}^K p(\theta^{1,i} | \gamma_1) p(\gamma_1 | \xi_1) \quad (26)$$

$$p(\gamma_2 | \theta^{2,1}, \dots, \theta^{2,K}) \propto \prod_{i=1}^K p(\theta^{2,i} | \gamma_2) p(\gamma_2 | \xi_2) \quad (27)$$

where  $p(\gamma_1 | \xi_1)$  and  $p(\gamma_2 | \xi_2)$  are vague priors over the hyperparameters, parameterised by  $\xi_1$  and  $\xi_2$ .

### 3.1.3 Posterior over the indicators

The conditional posterior distribution over the indicators is given by:

$$p(c_n = k | \mathbf{c}_{-n}, \mathbf{y}_n, \theta^k) \propto p(\mathbf{y}_n | \theta^k, c_n = k) p(c_n = k | \mathbf{c}_{-n}, \alpha_0) \quad (28)$$

## 3.2 Graphical model

The complete graphical model for the Dirichlet process mixture model of probabilistic CCA is shown in Figure 3, illustrating the layered structure of the hierarchical priors. Each pair of data observations  $\mathbf{y}_n = \{\mathbf{y}_{1,n}, \mathbf{y}_{2,n}\}$  is generated from one of the  $K$  represented pairs of mixture components, which is indicated by  $c_n$ . Each pair of mixture components is governed by a set of parameters, where the  $k$ th component pair's parameters are  $\theta^{1,k} = \{\mu_{1,k}, \mathbf{A}_{1,k}, \mathbf{W}_{1,k}\}$  and  $\theta^{2,k} = \{\mu_{2,k}, \mathbf{A}_{2,k}, \mathbf{W}_{2,k}\}$ . The parameter sets are governed by a set of hyperparameters  $\gamma_1$  and  $\gamma_2$ , which in turn are governed by vague priors  $\xi_1$  and  $\xi_2$ . The model and a Gibbs sampling scheme is derived in the next section in detail.

### 3.3 Priors over the component parameters

#### 3.3.1 Mean vector $\mu_k$

The mean vector for the  $k$ th latent variable model is drawn from a Gaussian distribution with hyperparameters  $\lambda$  and  $\mathbf{R}$  which are common to all components.

$$\mu_{1,k} \sim \mathcal{N}(\mu_{1,k} \mid \lambda_1, \mathbf{R}_1^{-1}) \quad (29)$$

$$\mu_{2,k} \sim \mathcal{N}(\mu_{2,k} \mid \lambda_2, \mathbf{R}_2^{-1}) \quad (30)$$

The hyperparameters  $\lambda_1, \lambda_2$  and  $\mathbf{R}_1, \mathbf{R}_2$  are given vague Normal and Wishart priors respectively,

$$\lambda_1 \sim \mathcal{N}(\lambda_1 \mid \mu_{y_1}, \Sigma_{y_1}), \mathbf{R}_1 \sim \mathcal{W}(\mathbf{R}_1 \mid m_1, \Sigma_{y_1}^{-1}) \quad (31)$$

$$\lambda_2 \sim \mathcal{N}(\lambda_2 \mid \mu_{y_2}, \Sigma_{y_2}), \mathbf{R}_2 \sim \mathcal{W}(\mathbf{R}_2 \mid m_2, \Sigma_{y_2}^{-1}) \quad (32)$$

where  $\mu_{y_1}$  and  $\Sigma_{y_1}$  are the sample mean and covariance of the first data set  $\mathbf{Y}_1$ , and  $\mu_{y_2}$  and  $\Sigma_{y_2}$  are the sample mean and covariance of the second data set  $\mathbf{Y}_2$ . The posterior distributions over the mean vectors are given by:

$$\mu_{1,k} \mid \theta_k, \mathbf{Y}_1 \sim \mathcal{N}(\mu_{1,k} \mid \mu_{\mu_{1,k}}, \Sigma_{\mu_{1,k}}) \quad (33)$$

where

$$\begin{aligned} \Sigma_{\mu_{1,k}} &= (N_k \Psi_{1,k}^{-1} + \mathbf{R}_1)^{-1} \\ \mu_{\mu_{1,k}} &= \Sigma_{\mu_{1,k}} (\Psi_{1,k}^{-1} N_k (\bar{\mathbf{y}}_{1,k} - \mathbf{W}_{1,k} \bar{\mathbf{x}}_k) + \mathbf{R}_1 \lambda_1) \\ \mu_{2,k} \mid \theta_k, \mathbf{Y}_2 &\sim \mathcal{N}(\mu_{2,k} \mid \mu_{\mu_{2,k}}, \Sigma_{\mu_{2,k}}) \quad (34) \end{aligned}$$

where

$$\begin{aligned} \Sigma_{\mu_{2,k}} &= (N_k \Psi_{2,k}^{-1} + \mathbf{R}_2)^{-1} \\ \mu_{\mu_{2,k}} &= \Sigma_{\mu_{2,k}} (\Psi_{2,k}^{-1} N_k (\bar{\mathbf{y}}_{2,k} - \mathbf{W}_{2,k} \bar{\mathbf{x}}_k) + \mathbf{R}_2 \lambda_2) \end{aligned}$$

where  $N_k$  is the number of data points in the  $k$ th cluster,  $\bar{\mathbf{y}}_{1,k} = \frac{1}{N_k} \sum_{i:c_i=k} \mathbf{y}_{1,i}$ ,  $\bar{\mathbf{y}}_{2,k} = \frac{1}{N_k} \sum_{i:c_i=k} \mathbf{y}_{2,i}$ ,  $\bar{\mathbf{x}}_k = \frac{1}{N_k} \sum_{i:c_i=k} \mathbf{x}_i$ .

The posterior distributions over the hyperparameters are given by:

$$\lambda_1 \mid \mu_{1,1}, \dots, \mu_{1,K}, \mathbf{R}_1 \sim \mathcal{N}\left(\lambda_1 \mid \frac{\mathbf{R}_1 \sum_{i=1}^K \mu_{1,i} + \Sigma_{y_1}^{-1} \mu_{y_1}}{K \mathbf{R}_1 + \Sigma_{y_1}^{-1}}, \frac{1}{K \mathbf{R}_1 + \Sigma_{y_1}^{-1}}\right) \quad (35)$$

$$\lambda_2 \mid \mu_{2,1}, \dots, \mu_{2,K}, \mathbf{R}_2 \sim \mathcal{N}\left(\lambda_2 \mid \frac{\mathbf{R}_2 \sum_{i=1}^K \mu_{2,i} + \Sigma_{y_2}^{-1} \mu_{y_2}}{K \mathbf{R}_2 + \Sigma_{y_2}^{-1}}, \frac{1}{K \mathbf{R}_2 + \Sigma_{y_2}^{-1}}\right) \quad (36)$$

$$\mathbf{R}_1 \mid \mu_{1,1}, \dots, \mu_{1,K}, \lambda_1 \sim \mathcal{W}\left(\mathbf{R}_1 \mid m_1 + K, \frac{m_1 + K}{\mathbf{S}_{\mu_1 + \Sigma_{y_1}/m_1}}\right) \quad (37)$$

$$\mathbf{R}_2 \mid \mu_{2,1}, \dots, \mu_{2,K}, \lambda_2 \sim \mathcal{W}\left(\mathbf{R}_2 \mid m_2 + K, \frac{m_2 + K}{\mathbf{S}_{\mu_2 + \Sigma_{y_2}/m_2}}\right) \quad (38)$$

where  $\mathbf{S}_{\mu_1} = \sum_{k=1}^K (\mu_{1,k} - \lambda_1)(\mu_{1,k} - \lambda_1)^\top$  and  $\mathbf{S}_{\mu_2} = \sum_{k=1}^K (\mu_{2,k} - \lambda_2)(\mu_{2,k} - \lambda_2)^\top$

#### 3.3.2 Covariance matrix $\Psi_{1,k}, \Psi_{2,k}$

We work with the inverse of  $\Psi_{1,k}$  and  $\Psi_{2,k}$ :  $\mathbf{A}_{1,k} = \Psi_{1,k}^{-1}$  and  $\mathbf{A}_{2,k} = \Psi_{2,k}^{-1}$ .  $\mathbf{A}_{1,k}$  and  $\mathbf{A}_{2,k}$  are drawn from Wishart distributions:

$$\mathbf{A}_{1,k} \sim \mathcal{W}(\mathbf{A}_{1,k} \mid \beta_1, \mathbf{C}_1^{-1}) \quad (39)$$

$$\mathbf{A}_{2,k} \sim \mathcal{W}(\mathbf{A}_{2,k} \mid \beta_2, \mathbf{C}_2^{-1}) \quad (40)$$

The hyperparameters  $\beta_1, \beta_2, \mathbf{C}_1$  and  $\mathbf{C}_2$  are common to all  $K$  components.  $(\beta_1 - m_1 + 1)$  and  $(\beta_2 - m_2 + 1)$  are given vague Gamma priors, and  $\mathbf{C}_1$  and  $\mathbf{C}_2$  are given vague Wishart priors:

$$\begin{aligned} (\beta_1 - m_1 + 1)^{-1} &\sim \mathcal{G}((\beta_1 - m_1 + 1)^{-1}, 1, 1) \quad (41) \end{aligned}$$

$$\begin{aligned} (\beta_2 - m_2 + 1)^{-1} &\sim \mathcal{G}((\beta_2 - m_2 + 1)^{-1}, 1, 1) \quad (42) \end{aligned}$$

$$\mathbf{C}_1 \sim \mathcal{W}(\mathbf{C}_1 \mid m_1, \Sigma_{y_1}) \quad (43)$$

$$\mathbf{C}_2 \sim \mathcal{W}(\mathbf{C}_2 \mid m_2, \Sigma_{y_2}) \quad (44)$$

The posterior distributions over the covariance matrices are given by:

$$\begin{aligned} \Psi_{1,k} \mid \theta_k, \mathbf{Y}_1 &\sim \mathcal{W}(\Psi_{1,k} \mid N_k + \beta_1, \frac{\mathbf{C}_1}{\beta_1} + \frac{N_k}{N_k + \beta_1} \mathbf{S}_{y_{1,k}}) \quad (45) \end{aligned}$$

$$\begin{aligned} \Psi_{2,k} \mid \theta_k, \mathbf{Y}_2 &\sim \mathcal{W}(\Psi_{2,k} \mid N_k + \beta_2, \frac{\mathbf{C}_2}{\beta_2} + \frac{N_k}{N_k + \beta_2} \mathbf{S}_{y_{2,k}}) \quad (46) \end{aligned}$$

where  $\mathbf{S}_{y_{1,k}} = \frac{1}{N_k} \sum_{i:c_i=k} (\mathbf{y}_{1,i} - \mu_{1,k})(\mathbf{y}_{1,i} - \mu_{1,k})^\top$ ,  $\mathbf{S}_{y_{2,k}} = \frac{1}{N_k} \sum_{i:c_i=k} (\mathbf{y}_{2,i} - \mu_{2,k})(\mathbf{y}_{2,i} - \mu_{2,k})^\top$ . The posterior distributions over the hyperparameters are given by:

$$\begin{aligned} \mathbf{C}_1 \mid \mathbf{A}_{1,1}, \dots, \mathbf{A}_{1,K}, \beta_1 &\sim \mathcal{W}(\mathbf{C}_1 \mid \beta_1 K + m_1, \frac{(\beta_1^{-1} \sum_{i=1}^K \mathbf{A}_{1,i} + m_1^{-1} \Sigma_{y_1}^{-1})^{-1}}{\beta_1 K + m_1}) \quad (47) \end{aligned}$$

$$\begin{aligned} \mathbf{C}_2 \mid \mathbf{A}_{2,1}, \dots, \mathbf{A}_{2,K}, \beta_2 &\sim \mathcal{W}(\mathbf{C}_2 \mid \beta_2 K + m_2, \frac{(\beta_2^{-1} \sum_{i=1}^K \mathbf{A}_{2,i} + m_2^{-1} \Sigma_{y_2}^{-1})^{-1}}{\beta_2 K + m_2}) \quad (48) \end{aligned}$$

and

$$\begin{aligned} \beta_1 \mid \mathbf{A}_{1,1}, \dots, \mathbf{A}_{1,K}, \mathbf{C}_1 &\propto \mathcal{G}((\beta_1 - m_1 + 1)^{-1}, 1, 1) \prod_k \mathcal{W}(\mathbf{A}_{1,k} \mid \beta_1, \mathbf{C}_1^{-1}) \quad (49) \end{aligned}$$

$$\begin{aligned} \beta_2 \mid \mathbf{A}_{2,1}, \dots, \mathbf{A}_{2,K}, \mathbf{C}_2 &\propto \mathcal{G}((\beta_2 - m_2 + 1)^{-1}, 1, 1) \prod_k \mathcal{W}(\mathbf{A}_{2,k} \mid \beta_2, \mathbf{C}_2^{-1}) \quad (50) \end{aligned}$$

Since the latter densities are not of standard form, independent samples are generated from  $\log \beta_1 \mid \mathbf{A}_{1,1}, \dots, \mathbf{A}_{1,K}$  and  $\log \beta_2 \mid \mathbf{A}_{2,1}, \dots, \mathbf{A}_{2,K}$  (which can be shown to be log concave distributions) using the Adaptive Rejection Sampling (ARS) technique [17].

#### 3.3.3 Weight vectors $\mathbf{W}_{1,k}, \mathbf{W}_{2,k}$

The weight matrices for the  $k$ th latent variable model are  $\mathbf{W}_{1,k}$  and  $\mathbf{W}_{2,k}$ . The rows of these matrices are drawn

from a Gaussian prior such that:

$$(\mathbf{W}_{1,k})^i \sim \mathcal{N}((\mathbf{W}_{1,k})^i \mid \gamma_{1,i}, v_1^{-1} \mathbf{I}_q) \quad (51)$$

$$(\mathbf{W}_{2,k})^i \sim \mathcal{N}((\mathbf{W}_{2,k})^i \mid \gamma_{2,i}, v_2^{-1} \mathbf{I}_q) \quad (52)$$

where  $(\mathbf{W}_{1,k})^i$  and  $(\mathbf{W}_{2,k})^i$  are the  $i$ th rows of  $\mathbf{W}_{1,k}$  and  $\mathbf{W}_{2,k}$  respectively,  $\gamma_{1,i}$  and  $\gamma_{2,i}$  are the means of the corresponding distributions, and  $v_1$  and  $v_2$  are the inverse variance. The hyperparameters are given the following vague priors:

$$v_1 \sim \mathcal{G}(v_1 \mid 1, 1) \quad (53)$$

$$v_2 \sim \mathcal{G}(v_2 \mid 1, 1) \quad (54)$$

$$\gamma_{1,i} \sim \mathcal{N}(\gamma_{1,i} \mid 0, \tau^{-1} \mathbf{I}_q) \quad (55)$$

$$\gamma_{2,i} \sim \mathcal{N}(\gamma_{2,i} \mid 0, \tau^{-1} \mathbf{I}_q) \quad (56)$$

The posterior distributions over the rows of the weight matrices are given by:

$$(\mathbf{W}_{1,k})^i \mid \theta_k, \mathbf{Y}_1 \sim \mathcal{N}((\mathbf{W}_{1,k})^i \mid \mu_{(\mathbf{W}_{1,k})^i}, \Sigma_{(\mathbf{W}_{1,k})^i}) \quad (57)$$

where

$$\Sigma_{(\mathbf{W}_{1,k})^i} = (\Psi_{1,k}(i,i))^{-1} \sum_{n:c_n=k} \mathbf{x}_n \mathbf{x}_n^\top + \mathbf{v}_1 \mathbf{I}_q)^{-1}$$

$$\mu_{(\mathbf{W}_{1,k})^i} = \frac{(\Sigma_{(\mathbf{W}_{1,k})^i})^{-1} \sum_{m:c_m=k} \mathbf{x}_m (\mathbf{y}_{1,m}(i) - \mu_{1,k}(i)) + \mathbf{v}_1 \gamma_{1,i}}{\Psi_{1,k}(i,i)}$$

$$(\mathbf{W}_{2,k})^i \mid \theta_k, \mathbf{Y}_2 \sim \mathcal{N}((\mathbf{W}_{2,k})^i \mid \mu_{(\mathbf{W}_{2,k})^i}, \Sigma_{(\mathbf{W}_{2,k})^i})$$

where

$$\Sigma_{(\mathbf{W}_{2,k})^i} = (\Psi_{2,k}(i,i))^{-1} \sum_{n:c_n=k} \mathbf{x}_n \mathbf{x}_n^\top + \mathbf{v}_2 \mathbf{I}_q)^{-1}$$

$$\mu_{(\mathbf{W}_{2,k})^i} = \frac{(\Sigma_{(\mathbf{W}_{2,k})^i})^{-1} \sum_{m:c_m=k} \mathbf{x}_m (\mathbf{y}_{2,m}(i) - \mu_{2,k}(i)) + \mathbf{v}_2 \gamma_{2,i}}{\Psi_{2,k}(i,i)}$$

The posterior distributions over the hyperparameters are given by:

$$\mathbf{v}_1 \mid \{\mathbf{W}_{1,k}\}_{k=1}^K, \{\gamma_{1,i}\}_{i=1}^{m_1} \sim \mathcal{G}(v_1 \mid \mu_{v_1}, s_{v_1}) \quad (58)$$

$$\mathbf{v}_2 \mid \{\mathbf{W}_{1,k}\}_{k=1}^K, \{\gamma_{1,i}\}_{i=1}^{m_1} \sim \mathcal{G}(v_2 \mid \mu_{v_2}, s_{v_2}) \quad (59)$$

where

$$\mu_{v_1} = m_1 K + 1$$

$$s_{v_1} = \frac{\mu_{v_1}}{1 + \sum_{i=1}^{m_1} \sum_k ((\mathbf{W}_{1,k})^i - \gamma_{1,i})^\top ((\mathbf{W}_{1,k})^i - \gamma_{1,i})}$$

$$\mu_{v_2} = m_2 K + 1$$

$$s_{v_2} = \frac{m_2 K + 1}{1 + \sum_{i=1}^{m_2} \sum_k ((\mathbf{W}_{2,k})^i - \gamma_{2,i})^\top ((\mathbf{W}_{2,k})^i - \gamma_{2,i})}$$

and

$$\gamma_{1,i} \mid \{(\mathbf{W}_{1,k})^i\}_{k=1}^K, v_1 \sim \mathcal{N}(\gamma_{1,i} \mid \frac{v_1 \sum_k (\mathbf{W}_{1,k})^i}{K v_1 + \tau}, \frac{1}{K v_1 + \tau}) \quad (60)$$

$$\gamma_{2,i} \mid \{(\mathbf{W}_{2,k})^i\}_{k=1}^K, v_2 \sim \mathcal{N}(\gamma_{2,i} \mid \frac{v_2 \sum_k (\mathbf{W}_{2,k})^i}{K v_2 + \tau}, \frac{1}{K v_2 + \tau}) \quad (61)$$

### 3.3.4 Latent variable $\mathbf{x}$

The latent variable  $\mathbf{x}_n$  for the  $n$ th pair of data points  $\mathbf{y}_n = [\mathbf{y}_{1,n}^\top, \mathbf{y}_{2,n}^\top]^\top$  is drawn from a Gaussian prior with zero mean and unit variance:

$$\mathbf{x}_n \sim \mathcal{N}(\mathbf{x}_n \mid 0, \mathbf{I}_q) \quad (62)$$

The posterior distribution over  $\mathbf{x}_n$  is given by:

$$\mathbf{x}_n \mid \theta, \mathbf{y}_{1,n}, \mathbf{y}_{2,n} \sim \mathcal{N}(\mathbf{x}_n \mid \mu_{\mathbf{x}_n}, \Sigma_{\mathbf{x}_n}) \quad (63)$$

where

$$\mu_{\mathbf{x}_n} = \mathbf{W}_{c_n}^\top (\mathbf{W}_{c_n} \mathbf{W}_{c_n}^\top + \Psi_{c_n})^{-1} (\mathbf{y}_n - \mu_{c_n})$$

$$\Sigma_{\mathbf{x}_n} = \mathbf{I}_q - \mathbf{W}_{c_n}^\top (\mathbf{W}_{c_n} \mathbf{W}_{c_n}^\top + \Psi_{c_n})^{-1} \mathbf{W}_{c_n}$$

where  $c_n \in \{1, \dots, K\}$  denotes the component index which generated  $\mathbf{y}_n$ ,  $\mathbf{W}_{c_n}$ ,  $\mu_{c_n}$  and  $\Psi_{c_n}$  are the parameters of the corresponding component, with  $\mathbf{W}_{c_n} = [\mathbf{W}_{1,c_n}^\top, \mathbf{W}_{2,c_n}^\top]^\top$ ,  $\mu_{c_n} = [\mu_{1,c_1}, \mu_{2,c_1}]^\top$ , and  $\Psi_{c_n} = \begin{pmatrix} \Psi_{1,c_n} & 0 \\ 0 & \Psi_{2,c_n} \end{pmatrix}$

### 3.3.5 Indicators $c_n$

The conditional priors on the indicators is given by:

$$c_n = k \mid \mathbf{c}_{-n}, \alpha_0 = \frac{N_{-n,k}}{N-1+\alpha_0} \quad (64)$$

$$c_n \neq c_{n'} \forall n' \neq n \mid \mathbf{c}_{-n}, \alpha = \frac{\alpha_0}{N-1+\alpha_0} \quad (65)$$

where  $-n$  indicates all the indices except  $n$ , such that  $\mathbf{c}_{-n}$  denotes all the indicators except the  $n$ th, and  $N_{-n,k}$  is the number of data points associated with the  $k$ th component, excluding the  $n$ th data point. A vague Gamma prior is placed over  $\alpha_0$ :

$$\alpha_0 \sim \mathcal{G}(\alpha_0 \mid 1, 1) \quad (66)$$

The posterior distributions over the indicators is given by the following:

for components for which  $N_{-n,k} > 0$

$$c_n = k \mid \mathbf{c}_{-n}, \theta_k, \alpha_0 \propto \frac{N_{-n,k}}{N-1+\alpha_0} \mathcal{N}(\mathbf{y}_n \mid \mathbf{W}_k \mathbf{x}_n + \mu_k, \Psi_k) \quad (67)$$

for all other components:

$$c_n \neq c_{n'} \forall n' \neq n \mid \mathbf{c}_{-n}, \gamma, \alpha_0 \propto \frac{\alpha_0}{N-1+\alpha_0} \int p(\mathbf{y}_n \mid \mathbf{x}_n, \theta^k) p(\mathbf{x}_n) p(\theta^k \mid \gamma) d\mathbf{x}_n d\theta^k \quad (68)$$

The likelihood for currently unrepresented classes (which have no parameters associated with them) is found by integrating over the parameter priors. The posterior distribution over  $\alpha_0$  is given by:

$$\alpha_0 \mid K, N \propto \frac{\alpha_0^K \Gamma(\alpha_0)}{\Gamma(\alpha_0 + N)} \mathcal{G}(\alpha_0 \mid 1, 1) \quad (69)$$

This only depends on  $N$  and  $K$ , and not on how the observations are distributed among the components. Samples are generated from  $p(\log(\alpha_0) \mid K, N)$ , which is log concave, using ARS.

### 3.4 Inference in the model

As noted before, exact analytical inference is not possible in this model, and Gibbs sampling is used to update the parameters, hyperparameters and indicator variables. Each variable in turn is updated by sampling from its posterior distribution conditional on all the other variables as follows:

- The parameters are updated by sampling from  $p(\theta \mid \gamma, \mathbf{c}, \mathbf{Y})$
- The hyperparameters are updated by sampling from  $p(\gamma \mid \theta, \mathbf{c}, \mathbf{Y})$
- The indicator variables are updated by sampling from  $p(\mathbf{c} \mid \theta, \gamma, \mathbf{Y})$
- The concentration parameter is updated by sampling from  $p(\log(\alpha_0) \mid K, N)$

This process (a Gibbs sweep) generates a sample from the joint posterior distribution  $p(\theta, \gamma, \mathbf{c} \mid \mathbf{Y})$ . Many Gibbs sweeps are performed to repeatedly update all the variables. Since consecutive samples are likely to be correlated, in order to generate independent samples from the joint posterior, the mixing time of the Markov chain is calculated and a sample is taken in every period of this length.

## 4 Experiments

To illustrate the model, we use a pair of toy data sets (each 2 dimensional) where the first data set follows an arc, and the second data set follows a sine curve, where the points correspond by arc length.

To perform inference for the model, we initialise the model with one component and then perform a large number of Gibbs sweeps to update the hyperparameters, parameters, and indicator variables, storing the values at each iteration. Initially, we do not know how the Markov chain will mix and converge for this particular data set so we perform 10000 iterations to assess the mixing and convergence times. The convergence time (or the burn-in time) is found to be approximately 3000 iterations. Discarding the 3000 iterations produced during the burn-in phase, the mixing time for the Markov chain is estimated by plotting the autocovariance for different parameters against time (based on 10000 iterations) and finding the maximum correlation length. There are no significant correlations for any of the parameters; we choose the effective correlation length to be 10 iterations. We then perform 10000 iterations for modelling purposes - 3000 for the burn-in period, and a further 7000 which generates 700 independent samples from the posterior distribution (spaced evenly 10 apart). Figure 6 shows four sets of samples from the posterior distribution for the mixture models at iterations 1, 500, 4000, and 6000. Figure 4 shows the histograms for some parameters of the

mixture model, based on the 700 independent samples from the posterior distribution.

### 4.1 Examining the distribution over the latent space

In the mixture model, there is a set of latent variables  $\mathbf{X}$  that underlies both data spaces  $\mathbf{Y}_1$  and  $\mathbf{Y}_2$ . In this section, we find the distribution over  $\mathbf{X}$  given just one of the data sets. This distribution can be used to predict one data set given the other, and vice versa. The posterior distribution over the  $n$ th latent variable  $\mathbf{x}_n$  given the corresponding data point from the first data set, is given by:

$$\begin{aligned} p(\mathbf{x}_n \mid \mathbf{y}_{1,n}) &= \int p(\mathbf{x}_n \mid \mathbf{y}_{1,n}, \theta) p(\theta) d\theta \quad (70) \\ &= \frac{1}{I} \sum_{i=1}^I \mathcal{N}(\mathbf{x}_n \mid (\mu_{\mathbf{x}_n})^i, (\Sigma_{\mathbf{x}_n})^i) \quad (71) \end{aligned}$$

where  $I$  is the number of independent samples, and the superscript  $i$  denotes the  $i$ th independent sample, such that  $\theta^i$  describes the  $i$ th sample of the posterior over  $\theta$ .

### 4.2 Predictive distribution

After finding the posterior distribution over the latent space given one data set, we can evaluate the predictive distribution over the other data space, according to:

$$\begin{aligned} p(\mathbf{y}_{2,n} \mid \mathbf{y}_{1,n}) \\ = \frac{1}{I} \sum_{i=1}^I \int p(\mathbf{y}_{2,n} \mid \mathbf{x}_n, \theta_i) p(\mathbf{x}_n \mid \mathbf{y}_{1,n}) d\mathbf{x}_n \quad (72) \end{aligned}$$

Figure 5 shows the predictive distribution over each data set given the other.

## 5 Conclusion

In this paper, we have presented a model to probabilistically represent the shared structure between two data sets. The model is a Dirichlet process mixture model of probabilistic canonical correlation analysers, allowing the complexity of the model (the number of mixture components) to be determined automatically from the data. This allows the model to automatically determine the flexibility of the manifolds underlying the data. The model offers a very flexible prior over the shared structure, overcoming the limitations of parametric Bayesian models and existing dependency seeking discriminative models. A future direction for research would be to use variational approximations to the model so that it would be possible to work with pairs of large scale data sets.

References:

[1] C. Fyfe and G. Leen. Stochastic Processes for Canonical Correlation Analysis. *European Symposium of Artificial Neural Networks (ESANN)*, 2006.

[2] H. Hotelling. Relations between two sets of variates. *Biometrika*, (28):312–377, 1936.

[3] M. Borga. *Learning multidimensional signal processing*. PhD thesis, Linköping University, Sweden, SE-583 83 Linköping, Sweden, Dissertation No. 531, ISBN 91-7219-202-X, 1998.

[4] P. L. Lai and C. Fyfe. Kernel and Nonlinear Canonical Correlation Analysis. *International Journal of Neural Systems*, 10(5):365–377, 2000.

[5] T. S. Ferguson. A Bayesian analysis of some non-parametric problems. *Annals of Statistics*, 2:209–230, 1973.

[6] F.R. Bach and M.I. Jordan. A Probabilistic Interpretation of Canonical Correlation Analysis. Technical Report 688, Dept of Statistics, University of California, 2005.

[7] M. Tipping and C. Bishop. Mixtures of probabilistic principal component analysers. Technical Report NCRG/97/003, Neural Computing Research Group, Aston University, 1997.

[8] D. Blackwell and J. B. MacQueen. Ferguson Distributions by Pólya Urn Schemes. In *Annals of Statistics*, number 1, pages 353–355. 1973.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39(1):1–38, 1977.

[10] R. M. Neal. Bayesian mixture modeling by Monte Carlo simulation. Technical Report CRG-TR-91-2, Department of Computer Science, University of Toronto, 1991.

[11] C.E. Rasmussen. The Infinite Gaussian Mixture Model. In *NIPS 12*, 2000.

[12] Z. Ghahramani and M. J. Beal. Variational inference for Bayesian mixtures of factor analyzers. In *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 2000.

[13] A. Corduneanu and C. M. Bishop. Variational Bayesian Model Selection for Mixture Distributions. In T. Jaakkola and T. Richardson, editors, *Artificial Intelligence and Statistics*, pages 27–34. Morgan Kaufmann, 2001.

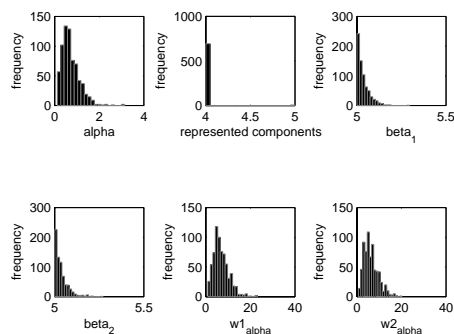


Figure 4: Some histograms for the posterior over different parameters in the model, given the data, based on 700 independent samples from the posterior

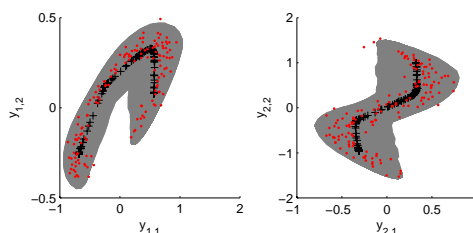


Figure 5: The predictive distribution over each data set given the other

[14] C. E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, 2:1152–1174, 1974.

[15] M. D. Escobar. Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, 89(425):268–277, 1994.

[16] M. West, P. Müller, and M. D. Escobar. Hierarchical priors and mixture models with applications in regression and density estimation. In P. R. Freeman and A. F. M Smith, editors, *Aspects of Uncertainty*, pages 363–386. 1994.

[17] W. R. Gilks and P. Wild. Adaptive rejection sampling for Gibbs sampling. In *Applied Statistics*, volume 41, pages 337–348. 1992.

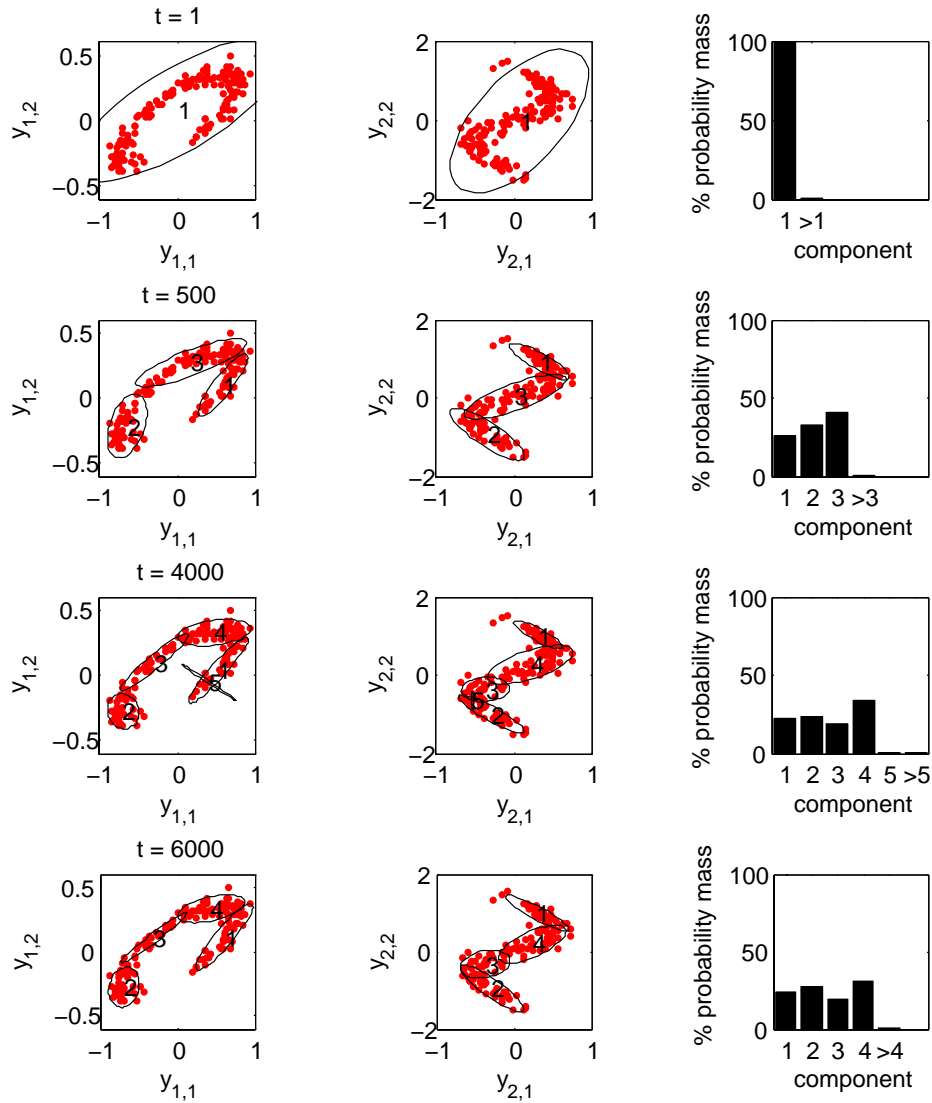


Figure 6: Four sets of samples from the posterior distribution for the mixture model, at iterations 1, 500, 4000, and 6000. Each row shows a sample over the first data set  $Y_1$  (first column) and the second data set  $Y_2$  (second column), and a graph for the probability mass in each component and the unrepresented components (third column). The ellipses indicate 2 standard deviations of the noise covariance matrices of each component, and the labels for each component 1,...,K are positioned at the means.