

The Hierarchical Local Partition Process

Lan Du

Minhua Chen

Qi An

Lawrence Carin

Department of Electrical and Computer Engineering

Duke University

Durham, NC 27708-0291, USA

LAN.DU@DUKE.EDU

MINHUA.CHEN@DUKE.EDU

QA@EE.DUKE.EDU

LCARIN@EE.DUKE.EDU

Aimee Zaas

Duke University Medical Center

Durham, NC 27708-0291, USA

AIMEE.ZAAS@DUKE.EDU

David B. Dunson

Department of Statistical Science

Duke University

Durham, NC 27708-0291, USA

DUNSON@STAT.DUKE.EDU

Editor:

Abstract

We consider the problem for which K different types of data are collected to characterize an associated inference task, with this performed for M distinct tasks. It is assumed that the parameters associated with the model for data type (modality) k may be represented in the form of a mixture model, with the M tasks representing M draws from the mixture. We wish to simultaneously infer mixture models across all K modality types, using data from all M tasks. Considering tasks m_1 and m_2 , we wish to impose the belief that if the data associated with modality k are drawn from the same mixture component (implying a similarity between tasks m_1 and m_2), then it is more probable that the associated data from modality $j \neq k$ will also be drawn from the same component. On the other hand, it is anticipated that there may be “random effects” that manifest idiosyncratic behavior for a subset of the modalities, even when similarity exists between the other modalities. The model employed utilizes a hierarchical Bayesian formalism, based on the local partition process. Inference is examined using both Markov chain Monte Carlo (MCMC) sampling and variational Bayesian (VB) analysis. The method is illustrated first with simulated data and then with data from two real applications. Concerning the latter, we consider analysis of gene-expression data and the sorting of annotated images.

Keywords: Clustering, Nonparametric Bayesian, Hierarchical Model, Multi-Task Learning, Dirichlet Process, Blocked Gibbs Sampling, Collapsed Gibbs Sampling, Variational Bayesian, Mixture Model, Hidden Cause, Random Effect, Sparse Factor Analysis, Data Fusion.

1. Introduction

Traditional methods of information retrieval, clustering or classification are typically based on one class of data. However, with the development of modern multimedia, a given item

may have joint representation in terms of words, images, audio, video, and other forms. It is of interest to develop techniques that process these often disparate types of data simultaneously, to infer inter-relationships between different information sources. While this framework is of interest in the modern world of ubiquitous multi-media information sources, it is also of general interest in multi-modality sensing. For example, in medicine one typically employs different tests and measurements to diagnose and treat an illness. The joint integration of these disparate data is of interest, as is the desire to examine the relationship between the multi-modality data from a given patient and a database of previous related data from other patients.

In the above examples the multi-modality data are observable. There are also problems that may be posed in terms of multiple latent modalities, or factors. An important example of this is Bayesian factor analysis (Carvalho et al.; Pournara and Wernish, 2007; Fokoue, 2004; Knowles and Ghahramani, 2007), and in this case the different “modalities” correspond to the latent factors. We again want to perform clustering of factors across multiple (*e.g.*, gene-expression) samples, but wish to do so in a flexible manner, as detailed further below.

A variety of methods from Bayesian statistics have been applied to multi-modality learning (Taskar et al., 2001; Jeon et al., 2003; Brochu et al., 2003; Barnard et al., 2003; Blei and Jordan, 2003; Wood et al., 2006; Airoidi et al., 2008; Kemp et al., 2004, 2006; Xu et al., 2006), with a focus on inferring the associations among the different types of data. In (Taskar et al., 2001; Jeon et al., 2003; Brochu et al., 2003), the relational data are considered as data pairs, and a generative model mixes the distinct data distributions in the relational feature space, to capture the probabilistic hidden causes. Authors (Barnard et al., 2003; Blei and Jordan, 2003) have also extended the simple mixture models in (Taskar et al., 2001; Brochu et al., 2003) by using the latent Dirichlet allocation (LDA) model (Blei et al., 2003) to allow the latent causes to be allocated repeatedly within a given annotated image. The Indian buffet process (IBP) prior (Griffiths and Ghahramani, 2005) has been used (Wood et al., 2006) to structure learning with infinite hidden causes. Recent extension of the latent stochastic blockmodel (Wang and Wong, 1987) applied mixture modeling based on a Dirichlet process prior for relational data, where each object belongs to a cluster and the relationships between objects are governed by the corresponding pair of clusters (Kemp et al., 2004, 2006; Xu et al., 2006). However, a limitation of such a model is that each object can only belong to one cluster. Since many relational data sets are multi-faceted (*i.e.* proteins or social actors), researchers (Airoidi et al., 2008) have relaxed the assumption of single-latent-role for actors, and developed a mixed membership model for relational data. The objective of integrating data from multiple modalities or “views” has also motivated non-Bayesian approaches. For example, the problem of interest here is also related to previous “multi-view” studies (Livescu et al., 2008), in which a given item is viewed from two or more distinct perspectives.

In this paper we consider K modalities associated with each “task”; a total of M tasks are considered. It is assumed that the data associated with each of the tasks represent draws from a mixture model. We wish to impose the belief that if data from modality k in tasks m_1 and m_2 are drawn from the same mixture, it is more probable that modality $j \neq k$ will also be drawn from a shared mixture model. This extends previous multi-task learning algorithms based upon the Dirichlet process (DP) (Xue et al., 2007), in which if sharing

existed between tasks m_1 and m_2 , then all model components across these two tasks were shared in the same way (termed here “global” sharing). At the other extreme, one may employ an independent DP to each of the K modalities, and in this case sharing, when it occurs, is only enforced “locally” within a given modality. The goal of mixing global and local sharing motivated the matrix stick-breaking process (MSBP) (Dunson et al., 2008). While the MSBP has proven successful in applications (Dunson et al., 2008), it leads to substantial computational difficulties as number of tasks and modalities increases, and is lack of interpretability.

This limitation of MSBP motivated development of the local partition process (LPP) prior for characterizing the joint distribution of multiple model parameters, with this performed within a Bayesian hierarchical model (Dunson). An LPP is constructed through a locally-weighted mixture of global and local clusterings, of which each clustering can be accomplished by a DP; therefore, it allows dependent local clustering and borrowing of information among different modalities (with respective model parameters). The original LPP model assumed that all data in a task are drawn from a single distribution. We are here interested in the case for which the data from a given task are assumed to be drawn from a mixture model; this motivates extending the original LPP by replacing the DP-type local and global components with hierarchical Dirichlet process (HDP) (Teh et al., 2006) construction; this new model is referred to as HLPP. A slice sampling (Walker, 2007) inference engine was developed for the original LPP model in (Dunson). However, such an MCMC inference algorithm converges slowly, especially for a large number of tasks. Therefore, additional contributions of this paper include development of a combination of collapsed (MacEachern, 1998) and blocked (Ishwaran and James, 2001) Gibbs sampling; we also develop a variational Bayesian (VB) (Beal, 2003) inference formulation for the HLPP model.

We demonstrate the proposed model by first using simulated data, followed by consideration of data from two real applications, *i.e.* gene-expression analysis and the analysis of annotated images. For the gene-expression analysis application, the HLPP prior is applied to a sparse factor analysis model (Carvalho et al.; Pournara and Wernish, 2007; Fokoue, 2004; Knowles and Ghahramani, 2007) to select the important factors and genes related to a disease; for the annotated-image application, the HLPP model is specialized to two modalities (image and text features, *i.e.* $K = 2$), with the goal of organizing/sorting multiple image-annotation pairs.

The remainder of the paper is organized as follows. In Section 2 we review the basic LPP model, and in Section 3 we extend this to an HLPP formulation. Section 4 develops the MCMC-based sampling scheme, as well as the VB inference formulation. A simulation example is presented in Section 5. The gene analysis and image-annotation applications are discussed in Section 6 and 7, respectively. Section 8 concludes the paper.

2. Local Partition Process

Assume M tasks for which we wish to infer models, and each task is characterized by K different feature types (“modalities”). Our goal is to *jointly* design mixture models for each of the K modalities, with the data in the M tasks representing M draws from the composite mixture model. We seek to learn models for all K modalities simultaneously, allowing

information from one modality to influence the mixture model for the other modalities. In the discussion that follows, for simplicity we assume the data from each of the K modalities are observable. In one of the real examples considered below (annotated images) the features are observable, while in the other example the “modalities” correspond to different sparse factors in a factor model (Carvalho et al.; Pournara and Wernish, 2007; Fokoue, 2004), and therefore the modalities are latent.

The data for task $m \in \{1, 2, \dots, M\}$ are $\mathbf{X}_m = \{\mathbf{x}_{lk}^{(m)}\}_{l=1, k=1}^{L_m, K}$, where L_m represents the number of samples in task m , and $\mathbf{x}_{lk}^{(m)}$ represents the l th sample in task m associated with modality k . It is assumed that when measuring the l th sample, data from all K modalities are acquired simultaneously; this is generalized when considering the text-image example in Section 7.4. Let $\mathbf{x}_{lk}^{(m)} \sim f_k(\boldsymbol{\theta}_{lk}^{(m)})$, for $k = 1, 2, \dots, K$; $m = 1, 2, \dots, M$; $l = 1, 2, \dots, L_m$; where $f_k(\cdot)$ is the likelihood for the k th modality, and $\boldsymbol{\theta}_{lk}^{(m)}$ represents the associated model parameters for the m th task, k th modality, and l th sample. We initially assume the same parameter $\boldsymbol{\theta}_k^{(m)}$ for all samples l , with this generalized in Section 3.

At one extreme, within the prior the parameters associated with the different modalities may be drawn independently, using independent DP priors for each modality:

$$\boldsymbol{\theta}_k^{(m)} \sim G_k, \quad G_k \sim \text{DP}(\alpha_k, H_k); \quad k = 1, 2, \dots, K; \quad m = 1, 2, \dots, M \quad (1)$$

This formulation does not impose (within the prior) statistical correlation between the modalities. As another extreme, one may consider a shared DP prior across all modalities:

$$\boldsymbol{\theta}^{(m)} \sim G, \quad G \sim \text{DP}(\alpha_0, \prod_{k=1}^K H_k); \quad m = 1, 2, \dots, M \quad (2)$$

where $\boldsymbol{\theta}_k^{(m)}$ represents the k th set of parameters associated with $\boldsymbol{\theta}^{(m)}$, and H_k denotes the base measure for the k th modality. The clustering properties being imposed via the priors in (1) and (2) may be understood by recalling that a draw $G \sim \text{DP}(\alpha, H)$ may be constructed as $G = \sum_{j=1}^{\infty} v_j \delta_{\boldsymbol{\Theta}_j}$, with $\mathbf{v} = \{v_j\}_{j=1}^{\infty}$ denoting the probability weights sampled from a stick-breaking process with parameter α , and $\delta_{\boldsymbol{\Theta}_j}$ denoting a probability measure concentrated at $\boldsymbol{\Theta}_j$ (Muliere and Tardella, 1998), with the $\boldsymbol{\Theta}_j$ drawn *i.i.d.* from H ; for notational convenience, we denote the draw of \mathbf{v} as $\mathbf{v} \sim \text{Stick}(\alpha)$. Hence, in (1) the clustering is imposed “locally” (independently) for each of the K modalities, while in (2) the clustering is imposed “globally” (simultaneously) across all K modalities. The “global” clustering shown in (2) is similar to the methods developed in (Taskar et al., 2001; Jeon et al., 2003; Brochu et al., 2003; Barnard et al., 2003; Blei and Jordan, 2003; Wood et al., 2006); the disadvantage of such an approach is that global sharing doesn’t account for random effects (for example) that may yield localized differences in some of the modalities. On the other hand, purely local sharing does not account for expected statistical correlations between the K modalities within a given task. We seek to impose the belief that if tasks m_1 and m_2 for modality k_1 are characterized by the same model parameters $\boldsymbol{\theta}_{k_1}^*$, then it is more likely that tasks m_1 and m_2 for modality k_2 will share the same parameter $\boldsymbol{\theta}_{k_2}^*$; however, the model should also allow random effects, in which some of the modalities from tasks m_1 and m_2 have distinct model parameters, despite the fact that many of the K modalities share parameters (in other words, within the prior we allow *partial* sharing of parameters

across the K modalities). We also desire that (1) and (2) are different limiting cases of our model.

To obtain a prior that addresses such a goal, (Petroni et al., 2008) proposed a hybrid functional DP that allows local allocation to clusters through a latent Gaussian process. Although the formulation is flexible, the use of a latent process to allow random effects in a subset of the modalities presents difficulties in the inference procedure. To avoid the complication of a latent Gaussian process, (Dunson) proposed a simpler model constructed through a weighted mixture of global and local clusterings (recall that “global” clustering occurs when all K modalities share parameters between tasks, and “local” clustering corresponds to independent clustering within a particular modality). The local partition process (LPP) is represented as

$$\begin{aligned}
\mathbf{x}_{lk}^{(m)} &\sim f_k(\boldsymbol{\theta}_k^{(m)}); \quad l = 1, 2, \dots, L_m; \quad k = 1, 2, \dots, K; \quad m = 1, 2, \dots, M \\
\begin{cases} \boldsymbol{\theta}_k^{(m)} = \boldsymbol{\vartheta}_k^{(m)} & \text{if } z_k^{(m)} = 0 \\ \boldsymbol{\theta}_k^{(m)} \sim G_k & \text{if } z_k^{(m)} = 1 \end{cases}; \quad k = 1, 2, \dots, K; \quad m = 1, 2, \dots, M \\
z_k^{(m)} &\sim \rho_k \delta_0 + (1 - \rho_k) \delta_1; \quad k = 1, 2, \dots, K; \quad m = 1, 2, \dots, M \\
\rho_k &\sim \text{Beta}(1, \beta_k); \quad k = 1, 2, \dots, K \\
\boldsymbol{\vartheta}^{(m)} &\sim G; \quad m = 1, 2, \dots, M \\
G &\sim \text{DP}(\alpha_0, \prod_{k=1}^K H_k) \\
G_k &\sim \text{DP}(\alpha_k, H_k); \quad k = 1, 2, \dots, K
\end{aligned} \tag{3}$$

where $\boldsymbol{\vartheta}^{(m)} = \{\boldsymbol{\vartheta}_k^{(m)}\}_{k=1}^K$; $z_k^{(m)} = i$ denotes the class of clustering associated with modality k in task m ($i = 0$ corresponds to global clustering and $i = 1$ to local clustering); ρ_k and $1 - \rho_k$ respectively denote the probabilities of global and local clusterings for modality k . Additionally, one may place gamma priors on the parameters $\boldsymbol{\alpha} = \{\alpha_{k'}\}_{k'=0}^K$ and $\boldsymbol{\beta} = \{\beta_k\}_{k=1}^K$. We typically favor β_k near zero such that ρ_k is likely to be near one, implying that most of the modalities are clustered in the same manner across the M tasks. However, with (relatively small) probability ρ_k , the k th modality will be clustered in an idiosyncratic manner, constituting “random effects”. To simplify notation below, we henceforth represent the 2nd to 6th lines in (3) as $\boldsymbol{\theta}_k^{(m)} \sim \text{LPP}(\boldsymbol{\beta}, G, \{G_k\}_{k=1}^K)$ for $k = 1, 2, \dots, K$ and $m = 1, 2, \dots, M$.

In the limit $\beta_k \rightarrow \infty$, we have $\rho_k = 0$, $z_k^{(m)} = 1$ and $\{\mathbf{x}_{lk}^{(m)}\}_{l=1}^{L_m}$ are generated through local (independent) clustering for $k = 1, 2, \dots, K$; in this case, we obtain (1). In the limit $\beta_k \rightarrow 0$, we have $\rho_k = 1$, $z_k^{(m)} = 0$ and $\{\mathbf{x}_{lk}^{(m)}\}_{l=1}^{L_m}$ are generated via global clustering; in this case, we obtain (2). For the general case $0 < \beta_k < \infty$ and hence $0 < \rho_k < 1$, the global clustering and local clusterings are combined by the mixing weight ρ_k and $1 - \rho_k$ for $k = 1, 2, \dots, K$. This model allows a greater degree of flexibility than the matrix stick-breaking process (MSBP) (Dunson et al., 2008), which was developed with the same basic goals (the MSBP model may achieve (1) as a limiting case, but not (2)).

Dunson proved the following clustering properties of the LPP model (Dunson):

$$\begin{aligned} \Pr(\boldsymbol{\theta}_{k_1}^{(m_1)} = \boldsymbol{\theta}_{k_1}^{(m_2)}) &= \left(\frac{1}{1+\alpha}\right) \left(\frac{1}{2+\beta}\right) \left(\beta + \frac{2}{1+\beta}\right) \\ \Pr(\boldsymbol{\theta}_{k_1}^{(m_1)} = \boldsymbol{\theta}_{k_1}^{(m_2)}, \boldsymbol{\theta}_{k_2}^{(m_1)} = \boldsymbol{\theta}_{k_2}^{(m_2)}) &= \left(\frac{1}{1+\alpha}\right)^2 \left(\frac{1}{2+\beta}\right)^2 \left[\left(\beta + \frac{2}{1+\beta}\right)^2 + \frac{4\alpha}{(1+\beta)^2} \right] \end{aligned} \quad (4)$$

Consequently, LPP has the properties:

- i) $\Pr(\boldsymbol{\theta}_{k_1}^{(m_1)} = \boldsymbol{\theta}_{k_1}^{(m_2)} | \boldsymbol{\theta}_{k_2}^{(m_1)} = \boldsymbol{\theta}_{k_2}^{(m_2)}) \geq \Pr(\boldsymbol{\theta}_{k_1}^{(m_1)} = \boldsymbol{\theta}_{k_1}^{(m_2)})$, for all $k_1 \neq k_2$ and $m_1 \neq m_2$.
- ii) $\Pr\left(\left[\Pr(\boldsymbol{\theta}_{k_1}^{(m_1)} = \boldsymbol{\theta}_{k_1}^{(m_2)} | \boldsymbol{\theta}_{k_2}^{(m_1)} = \boldsymbol{\theta}_{k_2}^{(m_2)}) - \Pr(\boldsymbol{\theta}_{k_1}^{(m_1)} = \boldsymbol{\theta}_{k_1}^{(m_2)})\right] \in S\right) > \varepsilon$, for any Borel subset $S \subset [0, 1]$, $k_1 \neq k_2$, $m_1 \neq m_2$, and for some $\varepsilon > 0$.

In the limit $\beta_k \rightarrow \infty$ for $k = 1, 2, \dots, K$, equality is achieved in Property i); while in the limit $\beta_k \rightarrow 0$ for $k = 1, 2, \dots, K$, $\Pr(\boldsymbol{\theta}_{k_1}^{(m_1)} = \boldsymbol{\theta}_{k_1}^{(m_2)} | \boldsymbol{\theta}_{k_2}^{(m_1)} = \boldsymbol{\theta}_{k_2}^{(m_2)}) = 1$. Therefore, when $0 < \beta_k < \infty$, if data $\{\mathbf{x}_{lk_2}^{(m_1)}\}_{l=1}^{L_{m_1}}$ and $\{\mathbf{x}_{lk_2}^{(m_2)}\}_{l=1}^{L_{m_2}}$ are contained within the same cluster, it is more probable that $\{\mathbf{x}_{lk_1}^{(m_1)}\}_{l=1}^{L_{m_1}}$ and $\{\mathbf{x}_{lk_1}^{(m_2)}\}_{l=1}^{L_{m_2}}$ will be in the same cluster. Property ii) implies that any degree of positive dependence in local clustering is supported by the LPP prior.

3. Integration of HDP and LPP

The LPP assumes all data in a task are drawn from a model with the same parameters, *i.e.* $\mathbf{x}_{lk}^{(m)} \sim f_k(\boldsymbol{\theta}_k^{(m)})$ for $l = 1, 2, \dots, L_k^{(m)}$. We now wish to extend this to a mixture model for $\boldsymbol{\theta}_{lk}^{(m)}$. To employ the LPP global and local clustering, with extension to a mixture model, we may substitute the DP-type local and global components of the original LPP with hierarchical Dirichlet processes (HDP) (Teh et al., 2006).

To construct an HDP, a probability measure $G_0 \sim \text{DP}(\gamma, H)$ is first drawn to define the base measure for each task, and then the measure associated with the m th task is $G^{(m)} \sim \text{DP}(\alpha, G_0)$; the parameters associated with the data in task m are drawn *i.i.d.* from $G^{(m)}$, therefore manifesting a mixture model within the task. Because G_0 is drawn from a DP, it is almost surely composed of a discrete set of atoms, and these atoms are shared across the M tasks, with different mixture weights (Teh et al., 2006). The HDP model is denoted as $\text{HDP}(\alpha, \gamma, H)$.

The hierarchical LPP (HLPP) is expressed as

$$\begin{aligned} \mathbf{x}_{lk}^{(m)} &\sim f_k(\boldsymbol{\theta}_{lk}^{(m)}), \quad \boldsymbol{\theta}_{lk}^{(m)} \sim \text{LPP}(\boldsymbol{\beta}, G^{(m)}, \{G_k^{(m)}\}_{k=1}^K); \\ & \quad l = 1, 2, \dots, L_m; \quad k = 1, 2, \dots, K; \quad m = 1, 2, \dots, M \\ G_k^{(m)} &\sim \text{HDP}(\alpha_k, \gamma_k, H_k); \quad k = 1, 2, \dots, K; \quad m = 1, 2, \dots, M \\ G^{(m)} &\sim \text{HDP}(\alpha_0, \gamma_0, \prod_{k=1}^K H_k); \quad m = 1, 2, \dots, M \end{aligned} \quad (5)$$

The global set of atoms are shared across the tasks with $z_k^{(m)} = 0$ for $k = 1, 2, \dots, K$, with the similarity in the global clustering controlled by the global task-specific atom weights;

the local set of atoms are shared across the tasks with $z_k^{(m)} = 1$ for modality k , with the similarity in the local clustering controlled by the local task-specific atom weights.

In the limit $\beta_k \rightarrow \infty$, we have $\rho_k = 0$, $z_k^{(m)} = 1$ for $k = 1, 2, \dots, K$, yielding

$$\begin{aligned}\boldsymbol{\theta}_{lk}^{(m)} &\sim G_k^{(m)}; \quad l = 1, 2, \dots, L_m; \quad k = 1, 2, \dots, K; \quad m = 1, 2, \dots, M \\ G_k^{(m)} &\sim \text{HDP}(\alpha_k, \gamma_k, H_k); \quad k = 1, 2, \dots, K; \quad m = 1, 2, \dots, M\end{aligned}\quad (6)$$

In this case the tasks are clustered locally (HDP is applied *independently* across the different modalities). In the limit $\beta_k \rightarrow 0$, we have $\rho_k = 1$, $z_k^{(m)} = 0$ for $k = 1, 2, \dots, K$, yielding

$$\begin{aligned}\boldsymbol{\theta}_l^{(m)} &\sim G^{(m)}; \quad l = 1, 2, \dots, L_m; \quad m = 1, 2, \dots, M \\ G^{(m)} &\sim \text{HDP}(\alpha_0, \gamma_0, \prod_{k=1}^K H_k); \quad m = 1, 2, \dots, M\end{aligned}\quad (7)$$

where $\boldsymbol{\theta}_l^{(m)} = \{\boldsymbol{\theta}_{lk}^{(m)}\}_{k=1}^K$. In this case all of the tasks are clustered globally (the HDP clustering is performed *simultaneously* across all K modalities). For the general case $0 < \beta_k < \infty$, $0 < \rho_k < 1$, similar to the original LPP model, the global clustering and local clusterings are combined by the mixing weight ρ_k and $1 - \rho_k$ for $k = 1, 2, \dots, K$.

We obtain similar clustering properties to those of the LPP model:

- i) $\Pr(\boldsymbol{\theta}_{l_1 k_1}^{(m_1)} = \boldsymbol{\theta}_{l_2 k_1}^{(m_2)} | \boldsymbol{\theta}_{l_3 k_2}^{(m_1)} = \boldsymbol{\theta}_{l_4 k_2}^{(m_2)}) \geq \Pr(\boldsymbol{\theta}_{l_1 k_1}^{(m_1)} = \boldsymbol{\theta}_{l_2 k_1}^{(m_2)})$, for all $k_1 \neq k_2$, $m_1 \neq m_2$, $l_1 \in \{1, 2, \dots, L_1^{(m_1)}\}$, $l_2 \in \{1, 2, \dots, L_1^{(m_2)}\}$, $l_3 \in \{1, 2, \dots, L_2^{(m_1)}\}$ and $l_4 \in \{1, 2, \dots, L_2^{(m_2)}\}$.
- ii) $\Pr\left(\left[\Pr(\boldsymbol{\theta}_{l_1 k_1}^{(m_1)} = \boldsymbol{\theta}_{l_2 k_1}^{(m_2)} | \boldsymbol{\theta}_{l_3 k_2}^{(m_1)} = \boldsymbol{\theta}_{l_4 k_2}^{(m_2)}) - \Pr(\boldsymbol{\theta}_{l_1 k_1}^{(m_1)} = \boldsymbol{\theta}_{l_2 k_1}^{(m_2)})\right] \in S\right) > \varepsilon$, for any Borel subset $S \subset [0, 1]$, $k_1 \neq k_2$, $m_1 \neq m_2$, $l_1 \in \{1, 2, \dots, L_1^{(m_1)}\}$, $l_2 \in \{1, 2, \dots, L_1^{(m_2)}\}$, $l_3 \in \{1, 2, \dots, L_2^{(m_1)}\}$, $l_4 \in \{1, 2, \dots, L_2^{(m_2)}\}$, and for some $\varepsilon > 0$.

The proofs are presented in Appendices A and B.

4. Posterior Computation

4.1 Stick-Breaking Construction of Draw from HDP

Assume that we wish to draw parameter $\boldsymbol{\theta}_l^{(m)}$ from $G^{(m)}$, where $G^{(m)} \sim \text{DP}(\alpha, G_0)$ and $G_0 \sim \text{DP}(\gamma, H)$. This may be represented via the stick-breaking construction as

$$\begin{aligned}\boldsymbol{\theta}_l^{(m)} &\sim \sum_{i=1}^{\infty} v_i^{(m)} \delta_{\Gamma_i^{(m)}}; \quad l = 1, 2, \dots, L_m; \quad m = 1, 2, \dots, M \\ \mathbf{v}^{(m)} &\sim \text{Stick}(\alpha); \quad m = 1, 2, \dots, M \\ \Gamma_i^{(m)} &\sim \sum_{j=1}^{\infty} \omega_j \delta_{\Theta_j}; \quad i = 1, 2, \dots, \infty; \quad m = 1, 2, \dots, M \\ \boldsymbol{\omega} &\sim \text{Stick}(\gamma) \\ \Theta_j &\sim H; \quad j = 1, 2, \dots, \infty\end{aligned}\quad (8)$$

Through the introduction of two indicator variables, the expression in (8) may now be expressed as

$$\begin{aligned}
\theta_l^{(m)} &= \Theta_{\zeta_l^{(m)}}; \quad l = 1, 2, \dots, L_m; \quad m = 1, 2, \dots, M \\
\xi_l^{(m)} &\sim \text{Mult}(\mathbf{v}^{(m)}); \quad l = 1, 2, \dots, L_m; \quad m = 1, 2, \dots, M \\
\mathbf{v}^{(m)} &\sim \text{Stick}(\alpha); \quad m = 1, 2, \dots, M \\
\zeta_g^{(m)} &\sim \text{Mult}(\omega); \quad g = 1, 2, \dots, \infty; \quad m = 1, 2, \dots, M \\
\omega &\sim \text{Stick}(\gamma) \\
\Theta_j &\sim H; \quad j = 1, 2, \dots, \infty
\end{aligned} \tag{9}$$

In (9) the indicator $\zeta_g^{(m)}$ represents which atom in the set $\{\Theta_j\}_{j=1}^\infty$ is associated with the g th stick in $G^{(m)}$, recognizing that the same atom may be used for multiple sticks. The indicator $\xi_l^{(m)}$ defines which atom from $G^{(m)}$ the l th sample in task m is drawn from. The advantage of (9) is that consecutive components in the hierarchy are in the conjugate-exponential family, aiding inference. The inference in (Teh et al., 2006) was performed using a Chinese-franchise construction, rather than a stick-breaking framework; the latter is useful for inference via the combination of collapsed (MacEachern, 1998) and blocked (Ishwaran and James, 2001) Gibbs sampling and VB inference (Beal, 2003).

The detailed HLPP model is

$$\begin{cases} \mathbf{x}_{lk}^{(m)} \sim f_k(\Theta_{k(0, \zeta_{0l}^{(m)})}^{(m)}) & \text{if } z_k^{(m)} = 0 \\ \mathbf{x}_{lk}^{(m)} \sim f_k(\Theta_{k(1, \zeta_{kl}^{(m)})}^{(m)}) & \text{if } z_k^{(m)} = 1 \end{cases}; \\
l = 1, 2, \dots, L_m; \quad k = 1, 2, \dots, K; \quad m = 1, 2, \dots, M \\
z_k^{(m)} \sim \rho_k \delta_0 + (1 - \rho_k) \delta_1; \quad k = 1, 2, \dots, K; \quad m = 1, 2, \dots, M \\
\rho_k \sim \text{Beta}(1, \beta_k); \quad k = 1, 2, \dots, K \\
\xi_{k'l}^{(m)} \sim \text{Mult}(\mathbf{v}_{k'}^{(m)}); \quad l = 1, 2, \dots, L_m; \quad k' = 0, 1, \dots, K; \quad m = 1, 2, \dots, M \\
\mathbf{v}_{k'}^{(m)} \sim \text{Stick}(\alpha_{k'}); \quad k' = 0, 1, \dots, K; \quad m = 1, 2, \dots, M \\
\zeta_{k'g}^{(m)} \sim \text{Mult}(\omega_{k'}); \quad k' = 0, 1, \dots, K; \quad m = 1, 2, \dots, M; \quad g = 1, 2, \dots, \infty \\
\omega_{k'} \sim \text{Stick}(\gamma_{k'}); \quad k' = 0, 1, \dots, K \\
\Theta_{k(i,j)} \sim H; \quad i = 0, 1; \quad j = 1, 2, \dots, \infty; \quad k = 1, 2, \dots, K
\end{aligned} \tag{10}$$

where $\{\Theta_{k(0,j)}\}_{j=1}^J$ and $\{\Theta_{k(1,j)}\}_{j=1}^J$ respectively represent the global and local set of atoms for modality k ; $v_{0j}^{(m)}$ is the j th component of the global task-specific atom weights $\mathbf{v}_0^{(m)}$, and $v_{kj}^{(m)}$ is the j th component of the local task-specific atom weights $\mathbf{v}_k^{(m)}$. The tasks that have similar global clustering will have similar $\mathbf{v}_0^{(m)}$; while the tasks that have similar local clustering for a given modality k will have similar $\mathbf{v}_k^{(m)}$. While the model may appear somewhat complicated, we note that the last five lines in (10) correspond to repeated application of the last five lines of the HDP representation in (9); $k = 0$ corresponds to

“global” HDP, and $k = 1, \dots, K$ correspond to the independent “local” HDPs on the respective K modalities. As before, the indicator $z_k^{(m)}$ defines whether the atoms associated with the k th modality in task m come from the “global” or “local” atoms.

For inference purposes, we truncate the number of sticks in $G^{(m)}$ and $G_k^{(m)}$ to T , and the number of sticks in G_0 and G_{0k} to J (see the truncation properties of the stick-breaking representation discussed in Muliere and Tardella, 1998); we impose $T \geq J$. Since $\boldsymbol{\alpha} = \{\alpha_{k'}\}_{k'=0}^K$, $\boldsymbol{\beta} = \{\beta_k\}_{k=1}^K$ and $\boldsymbol{\gamma} = \{\gamma_{k'}\}_{k'=0}^K$ are key hyper-parameters controlling the probability of global and local clusterings, we also place Gamma hyper-priors on them to allow the data to inform about their values, *i.e.* $\alpha_{k'} \sim \text{Ga}(a_\alpha, b_\alpha)$, $\beta_k \sim \text{Ga}(a_\beta, b_\beta)$ and $\gamma_{k'} \sim \text{Ga}(a_\gamma, b_\gamma)$.

To save space, we only present the Gibbs sampling procedure and VB update equations for variables that are unique to the current analysis; the reader is referred to (Teh et al., 2006) and (An et al., 2008) for related expressions employed within the HDP model.

4.2 Combination of Collapsed and Blocked Gibbs Sampling

We follow Bayes’ rule to derive the full conditional distribution for each random variable in the posterior distribution

$$p(\boldsymbol{\Phi}|\mathbf{X}, \boldsymbol{\Psi}) = \frac{p(\mathbf{X}|\boldsymbol{\Phi})p(\boldsymbol{\Phi}|\boldsymbol{\Psi})}{\int p(\mathbf{X}|\boldsymbol{\Phi})p(\boldsymbol{\Phi}|\boldsymbol{\Psi}) d\boldsymbol{\Phi}} \quad (11)$$

where $\boldsymbol{\Phi} = \left\{ \{\boldsymbol{\Theta}_{k(i,j)}\}, \{\boldsymbol{\omega}_{k'}\}, \{\mathbf{v}_{k'}^{(m)}\}, \{\rho_k\}, \{\zeta_{k'g}^{(m)}\}, \{\xi_{k'l}^{(m)}\}, \{z_k^{(m)}\} \right\}$ are hidden variables of interest, $\mathbf{X} = \{\mathbf{x}_{lk}^{(m)}\}$ are observed variables and $\boldsymbol{\Psi} = \{\{\alpha_{k'}\}, \{\beta_k\}, \{\gamma_{k'}\}, \boldsymbol{\Omega}\}$ are hyper-parameters ($\boldsymbol{\Omega}$ denotes the hyper-parameters of $H_k(\boldsymbol{\Theta}_{k(i,j)})$). The conditional posterior computation proceeds through the following steps:

- Sample $\boldsymbol{\Theta}_{k(i,j)}$ for $i = 0, 1$; $k = 1, 2, \dots, K$ and $j = 1, 2, \dots, J$ from

$$\begin{aligned} p(\boldsymbol{\Theta}_{k(0,j)}|\dots) &\propto H_k(\boldsymbol{\Theta}_{k(0,j)}) \left[\prod_{m,l:z_k^{(m)}=0, \zeta_{0l}^{(m)}=j} f_k(\mathbf{x}_{lk}^{(m)}; \boldsymbol{\Theta}_{k(0,j)}) \right]; \\ p(\boldsymbol{\Theta}_{k(1,j)}|\dots) &\propto H_k(\boldsymbol{\Theta}_{k(1,j)}) \left[\prod_{m,l:z_k^{(m)}=1, \zeta_{kl}^{(m)}=j} f_k(\mathbf{x}_{lk}^{(m)}; \boldsymbol{\Theta}_{k(1,j)}) \right] \end{aligned} \quad (12)$$

Note that if no observation is assigned to a specific cluster, then the parameters are drawn from the prior distribution $H_k(\boldsymbol{\Theta}_{k(i,j)})$. Also, if the prior is conjugate to the likelihood then sampling is greatly simplified. However, non-conjugate priors can be accommodated using rejection sampling or Metropolis-Hastings steps.

- Sample $\omega_{k'j}$ for $k' = 0, 1, \dots, K$ and $j = 1, 2, \dots, J$ by generating

$$\begin{aligned}
p(\omega'_{k'j} | \dots) &\propto \\
&\text{Beta} \left(1 + \sum_{m=1}^M \sum_{g=1}^T \mathbf{1}(\zeta_{k'g}^{(m)} = j), \gamma_{k'} + \sum_{m=1}^M \sum_{g=1}^T \sum_{h:h>j}^J \mathbf{1}(\zeta_{k'g}^{(m)} = h) \right); \\
& \quad \quad \quad j = 1, 2, \dots, J-1; \\
\omega'_{k'J} &= 1
\end{aligned} \tag{13}$$

and constructing $\omega_{k'j} = \omega'_{k'j} \prod_{h=1}^{j-1} (1 - \omega'_{k'h})$.

- Sample $v_{k'g}^{(m)}$ for $k' = 0, 1, \dots, K$; $g = 1, 2, \dots, T$ and $m = 1, 2, \dots, M$ by generating

$$\begin{aligned}
p(v_{k'g}^{(m)} | \dots) &\propto \text{Beta} \left(1 + \sum_{l=1}^{L_m} \mathbf{1}(\xi_{k'l}^{(m)} = g), \alpha_{k'} + \sum_{l=1}^{L_m} \sum_{h:h>g}^T \mathbf{1}(\xi_{k'l}^{(m)} = h) \right); \\
& \quad \quad \quad g = 1, 2, \dots, T-1; \\
v_{k'T}^{(m)} &= 1
\end{aligned} \tag{14}$$

and constructing $v_{k'g}^{(m)} = v_{k'g}^{(m)} \prod_{h=1}^{g-1} (1 - v_{k'h}^{(m)})$.

- Sample ρ_k for $k = 1, 2, \dots, K$ from

$$p(\rho_k | \dots) \propto \text{Beta} \left(1 + \sum_{m=1}^M \mathbf{1}(z_k^{(m)} = 0), \beta_k + \sum_{m=1}^M \mathbf{1}(z_k^{(m)} = 1) \right) \tag{15}$$

- Sample $\zeta_{k'g}^{(m)}$ for $k' = 0, 1, \dots, K$; $g = 1, 2, \dots, T$ and $m = 1, 2, \dots, M$ from a multinomial distribution with probabilities

$$\begin{aligned}
p(\zeta_{0g}^{(m)} = j | \dots) &\propto p(\zeta_{0g}^{(m)} = j | \{\zeta_{0g'}^{(m)}\}_{g' \neq g}, \gamma_0) \prod_{k:z_k^{(m)}=0, l:\xi_{0l}^{(m)}=g} f_k(x_{lk}^{(m)}; \Theta_{k(0,j)}) \\
&= \left(\frac{1 + \sum_{g' \neq g} \mathbf{1}(\zeta_{0g'}^{(m)} = j)}{1 + \gamma_0 + \sum_{g' \neq g} \mathbf{1}(\zeta_{0g'}^{(m)} \geq j)} \prod_{h < j} \frac{\gamma_0 + \sum_{g' \neq g} \mathbf{1}(\zeta_{0g'}^{(m)} > h)}{1 + \gamma_0 + \sum_{g' \neq g} \mathbf{1}(\zeta_{0g'}^{(m)} \geq h)} \right) \quad ; \\
& \quad \quad \quad \prod_{k:z_k^{(m)}=0, l:\xi_{0l}^{(m)}=g} f_k(x_{lk}^{(m)}; \Theta_{k(0,j)}) \\
p(\zeta_{kg}^{(m)} = j | \dots) &\propto p(\zeta_{kg}^{(m)} = j | \{\zeta_{kg'}^{(m)}\}_{g' \neq g}, \gamma_k) \mathbf{1}(z_k^{(m)} = 1) \prod_{l:\xi_{kl}^{(m)}=g} f_k(x_{lk}^{(m)}; \Theta_{k(1,j)}) \\
&= \left(\frac{1 + \sum_{g' \neq g} \mathbf{1}(\zeta_{kg'}^{(m)} = j)}{1 + \gamma_k + \sum_{g' \neq g} \mathbf{1}(\zeta_{kg'}^{(m)} \geq j)} \prod_{h < j} \frac{\gamma_k + \sum_{g' \neq g} \mathbf{1}(\zeta_{kg'}^{(m)} > h)}{1 + \gamma_k + \sum_{g' \neq g} \mathbf{1}(\zeta_{kg'}^{(m)} \geq h)} \right) \tag{16} \\
& \quad \quad \quad \mathbf{1}(z_k^{(m)} = 1) \prod_{l:\xi_{kl}^{(m)}=g} f_k(x_{lk}^{(m)}; \Theta_{k(1,j)})
\end{aligned}$$

- Sample $\xi_{k'l}^{(m)}$ for $k' = 0, 1, \dots, K$; $l = 1, 2, \dots, L_m$ and $m = 1, 2, \dots, M$ from a multinomial distribution with probabilities

$$\begin{aligned}
p(\xi_{0l}^{(m)} = g | \dots) &\propto p(\xi_{0l}^{(m)} = g | \{\xi_{0l'}^{(m)}\}_{l' \neq l}, \alpha_0) \prod_{k: z_k^{(m)}=0} f_k(x_{lk}^{(m)}; \Theta_{k(0, \zeta_{0g}^{(m)})}) \\
&= \left(\frac{1 + \sum_{l' \neq l} \mathbf{1}(\xi_{0l'}^{(m)} = g)}{1 + \alpha_0 + \sum_{l' \neq l} \mathbf{1}(\xi_{0l'}^{(m)} \geq g)} \prod_{h < g} \frac{\alpha_0 + \sum_{l' \neq l} \mathbf{1}(\xi_{0l'}^{(m)} > h)}{1 + \alpha_0 + \sum_{l' \neq l} \mathbf{1}(\xi_{0l'}^{(m)} \geq h)} \right) ; \\
&\quad \prod_{k: z_k^{(m)}=0} f_k(x_{lk}^{(m)}; \Theta_{k(0, \zeta_{0g}^{(m)})}) \\
p(\xi_{kl}^{(m)} = g | \dots) &\propto p(\xi_{kl}^{(m)} = g | \{\xi_{kl'}^{(m)}\}_{l' \neq l}, \alpha_k) \mathbf{1}(z_k^{(m)} = 1) f_k(x_{lk}^{(m)}; \Theta_{k(1, \zeta_{kg}^{(m)})}) \\
&= \left(\frac{1 + \sum_{l' \neq l} \mathbf{1}(\xi_{kl'}^{(m)} = g)}{1 + \alpha_k + \sum_{l' \neq l} \mathbf{1}(\xi_{kl'}^{(m)} \geq g)} \prod_{h < g} \frac{\alpha_k + \sum_{l' \neq l} \mathbf{1}(\xi_{kl'}^{(m)} > h)}{1 + \alpha_k + \sum_{l' \neq l} \mathbf{1}(\xi_{kl'}^{(m)} \geq h)} \right) \quad (17) \\
&\quad \mathbf{1}(z_k^{(m)} = 1) f_k(x_{lk}^{(m)}; \Theta_{k(1, \zeta_{kg}^{(m)})})
\end{aligned}$$

- Sample $z_k^{(m)}$ for $k = 1, 2, \dots, K$, and $m = 1, 2, \dots, M$ from a Bernoulli distribution with probabilities

$$\begin{aligned}
p(z_k^{(m)} = 0 | \dots) &\propto p(z_k^{(m)} = 0 | \{z_{k''}^{(m)}\}_{k'' \neq k}, \beta_k) \prod_{l=1}^{L_m} f_k(x_{lk}^{(m)}; \Theta_{k(0, \zeta_{0l}^{(m)})}) \\
&= \frac{1 + \sum_{k'' \neq k} \mathbf{1}(z_{k''}^{(m)} = 0)}{1 + \beta_k + \sum_{k'' \neq k} \mathbf{1}(z_{k''}^{(m)} \geq 0)} \prod_{l=1}^{L_m} f_k(x_{lk}^{(m)}; \Theta_{k(0, \zeta_{0l}^{(m)})}) ; \\
p(z_k^{(m)} = 1 | \dots) &\propto p(z_k^{(m)} = 1 | \{z_{k''}^{(m)}\}_{k'' \neq k}, \beta_k) \prod_{l=1}^{L_m} f_k(x_{lk}^{(m)}; \Theta_{k(1, \zeta_{kl}^{(m)})}) \\
&= \left(\frac{1 + \sum_{k'' \neq k} \mathbf{1}(z_{k''}^{(m)} = 1)}{1 + \beta_k + \sum_{k'' \neq k} \mathbf{1}(z_{k''}^{(m)} \geq 1)} \frac{\beta_k + \sum_{k'' \neq k} \mathbf{1}(z_{k''}^{(m)} > 0)}{1 + \beta_k + \sum_{k'' \neq k} \mathbf{1}(z_{k''}^{(m)} \geq 0)} \right) \quad (18) \\
&\quad \prod_{l=1}^{L_m} f_k(x_{lk}^{(m)}; \Theta_{k(1, \zeta_{kl}^{(m)})})
\end{aligned}$$

In the above sampling procedure, we use collapsed Gibbs sampling for the three indicator variables $\zeta_{k'g}^{(m)}$, $\xi_{k'l}^{(m)}$ and $z_k^{(m)}$, since the corresponding prior variables $\omega_{k'}$, $\mathbf{v}_{k'}$ and ρ can be marginalized out. For example, if we use blocked Gibbs sampling, (16) should be

$$\begin{aligned}
p(\zeta_{0g}^{(m)} = j | \dots) &\propto \omega_{0j} \prod_{k: z_k^{(m)}=0, l: \xi_{0l}^{(m)}=g} f_k(x_{lk}^{(m)}; \Theta_{k(0, j)}); \\
p(\zeta_{kg}^{(m)} = j | \dots) &\propto \omega_{kj} \mathbf{1}(z_k^{(m)} = 1) \prod_{l: \xi_{kl}^{(m)}=g} f_k(x_{lk}^{(m)}; \Theta_{k(1, j)}) \quad (19)
\end{aligned}$$

In this general case, collapsed Gibbs sampling improves upon blocked Gibbs sampling by marginalizing out the stick variables for the indicators, therefore dealing with them more exactly. As discussed in (MacEachern, 1998), if $f_k(\cdot)$ is a multinomial distribution and its prior is a Dirichlet distribution, we can further marginalize out $\Theta_{k(i, j)}$. However, here we just consider the general case.

The full posterior $p(\Phi | \mathbf{X}, \Psi)$ can be constructed by collecting a sufficient number of samples after the above iteration stabilizes (Gilks and Spiegelhalter, 1996). In our analysis of the MCMC sampler, we have considered and confirmed the convergence tests as described in (Geweke, 1992) and (Raftery and Lewis, 1992).

4.3 Variational Bayesian (VB) Inference

Although MCMC typically yields accurate results (with sufficient samples) (Gilks and Spiegelhalter, 1996), it often requires significant computational resources and the convergence of the algorithm is often difficult to diagnose. Variational Bayes (VB) inference (Beal, 2003) is an alternative method for approximating likelihoods and posteriors. Instead of directly estimating $p(\Phi|\mathbf{X}, \Psi)$, variational methods seek a distribution $q(\Phi)$ to approximate the true posterior distribution $p(\Phi|\mathbf{X}, \Psi)$. Considering the log “marginal likelihood”, *i.e.* the integration in the denominator of (11)

$$\log p(\mathbf{X}|\Psi) = \mathcal{L}(q(\Phi)) + \mathcal{D}_{KL}(q(\Phi)||p(\Phi|\mathbf{X}, \Psi)) \quad (20)$$

where

$$\mathcal{L}(q(\Phi)) = \int q(\Phi) \log \frac{p(\mathbf{X}|\Phi)p(\Phi|\Psi)}{q(\Phi)} d\Phi \quad (21)$$

and

$$\mathcal{D}_{KL}(q(\Phi)||p(\Phi|\mathbf{X}, \Psi)) = \int q(\Phi) \log \frac{q(\Phi)}{p(\Phi|\mathbf{X}, \Psi)} d\Phi \quad (22)$$

Since the Kullback-Leibler (KL) divergence between the approximation $q(\Phi)$ and true posterior $p(\Phi|\mathbf{X}, \Psi)$ denoted by $\mathcal{D}_{KL}(q(\Phi)||p(\Phi|\mathbf{X}, \Psi))$ is nonnegative, from (20) the approximation of the true posterior $p(\Phi|\mathbf{X}, \Psi)$ using $q(\Phi)$ can be achieved by maximization of $\mathcal{L}(q(\Phi))$, which forms a strict lower bound on $\log p(\mathbf{X}|\Psi)$,

$$\log p(\mathbf{X}|\Psi) \geq \mathcal{L}(q(\Phi)) \quad (23)$$

For computational convenience, $q(\Phi)$ is expressed in a factorized form, with the same functional form as the priors $p(\Phi|\Psi)$ and each parameter is represented by its own conjugate prior. For the HLPP model proposed in this paper, we assume

$$\begin{aligned} q(\Phi) &= q\left(\{\Theta_{k(i,j)}\}, \{\omega_{k'}\}, \{\mathbf{v}_{k'}^{(m)}\}, \{\rho_k\}, \{\zeta_{k'g}^{(m)}\}, \{\xi_{k'l}^{(m)}\}, \{z_k^{(m)}\}\right) \\ &= \left[\prod_{k=1}^K \prod_{i=0}^1 \prod_{j=1}^J q(\Theta_{k(i,j)}) \right] \left[\prod_{k'=0}^K q(\omega_{k'}) \right] \left[\prod_{m=1}^M \prod_{k'=0}^K q(\mathbf{v}_{k'}^{(m)}) \right] \left[\prod_{k=1}^K q(\rho_k) \right] \\ &\quad \cdot \left[\prod_{m=1}^M \prod_{k'=0}^K \prod_{g=1}^T q(\zeta_{k'g}^{(m)}) \right] \left[\prod_{k'=0}^K \prod_{m=1}^M \prod_{l=1}^{L_m} q(\xi_{k'l}^{(m)}) \right] \left[\prod_{k=1}^K \prod_{m=1}^M q(z_k^{(m)}) \right] \end{aligned} \quad (24)$$

In the VB algorithm, the maximization of the lower bound in (21) is realized by taking functional derivatives with respect to each term in (24), while fixing the other $q(\cdot)$ distributions and setting $\partial\mathcal{L}(q)/\partial q(\cdot) = 0$ (Beal, 2003). The update equations for the variational posterior are listed as follows.

- Update $\Theta_{k(i,j)}$ for $i = 0, 1$; $k = 1, 2, \dots, K$ and $j = 1, 2, \dots, J$ according to

$$\begin{aligned} \log q(\Theta_{k(0,j)}) &\propto \log H_k(\Theta_{k(0,j)}) + \left[\frac{\sum_{m=1}^M \sum_{k=1}^K q(z_k^{(m)} = 0) \sum_{g=1}^T q(\zeta_{0g}^{(m)} = j)}{\sum_{l=1}^{L_m} q(\xi_{0l}^{(m)} = g) \log f_k(\mathbf{x}_{lk}^{(m)}; \Theta_{k(0,j)})} \right]; \\ \log q(\Theta_{k(1,j)}) &\propto \log H_k(\Theta_{k(1,j)}) + \left[\frac{\sum_{m=1}^M \sum_{k=1}^K q(z_k^{(m)} = 1) \sum_{g=1}^T q(\zeta_{kg}^{(m)} = j)}{\sum_{l=1}^{L_m} q(\xi_{kl}^{(m)} = g) \log f_k(\mathbf{x}_{lk}^{(m)}; \Theta_{k(1,j)})} \right] \end{aligned} \quad (25)$$

If the prior is conjugate to the likelihood then we can easily get the update equations.

- Update $\omega'_{k'j}$ for $k' = 0, 1, \dots, K$ and $j = 1, 2, \dots, J - 1$ according to

$$q(\omega'_{k'j}) = \text{Beta}(\omega'_{k'j}; \tilde{\gamma}_{k'j}^{(1)}, \tilde{\gamma}_{k'j}^{(2)})$$

$$\tilde{\gamma}_{k'j}^{(1)} = 1 + \sum_{m=1}^M \sum_{g=1}^T q(\zeta_{k'g}^{(m)} = j), \quad \tilde{\gamma}_{k'j}^{(2)} = \gamma_{k'} + \sum_{m=1}^M \sum_{g=1}^T \sum_{h:h>j}^J q(\zeta_{k'g}^{(m)} = h) \quad (26)$$

Since $\omega_{k'j} = \omega'_{k'j} \prod_{h=1}^{j-1} (1 - \omega'_{k'h})$, updating of $\omega_{k'}$ is equivalent to updating the posterior of $\omega_{k'}$.

- Update $v_{k'g}^{(m)}$ for $k' = 0, 1, \dots, K$; $g = 1, 2, \dots, T$ and $m = 1, 2, \dots, M$ according to

$$q(v_{k'g}^{(m)}) = \text{Beta}(v_{k'g}^{(m)}; \tilde{\alpha}_{k'g}^{(m,1)}, \tilde{\alpha}_{k'g}^{(m,2)});$$

$$\tilde{\alpha}_{k'g}^{(m,1)} = 1 + \sum_{l=1}^{L_m} q(\xi_{k'l}^{(m)} = g), \quad \tilde{\alpha}_{k'g}^{(m,2)} = \alpha_{k'} + \sum_{l=1}^{L_m} \sum_{h:h>g}^T q(\xi_{k'l}^{(m)} = h) \quad (27)$$

Since $v_{k'g}^{(m)} = v_{k'g}^{(m)} \prod_{h=1}^{g-1} (1 - v_{k'h}^{(m)})$, updating of $v_{k'}$ is equivalent to updating the posterior of $v_{k'}$.

- Update ρ_k for $k = 1, 2, \dots, K$ according to

$$q(\rho_k) = \text{Beta}(\rho_k; \tilde{\beta}_{k1}, \tilde{\beta}_{k2});$$

$$\tilde{\beta}_{k1} = 1 + \sum_{m=1}^M q(z_k^{(m)} = 0), \quad \tilde{\beta}_{k2} = \beta_k + \sum_{m=1}^M q(z_k^{(m)} = 1) \quad (28)$$

- Update $\zeta_{k'g}^{(m)}$ for $k' = 0, 1, \dots, K$; $g = 1, 2, \dots, T$ and $m = 1, 2, \dots, M$ according to a multinomial distribution with probabilities

$$q(\zeta_{0g}^{(m)} = j) \propto \exp \left(\begin{aligned} & \sum_{h=1}^{j-1} [\psi(\tilde{\gamma}_{0h}^{(2)}) - \psi(\tilde{\gamma}_{0h}^{(1)} + \tilde{\gamma}_{0h}^{(2)})] + [\psi(\tilde{\gamma}_{0j}^{(1)}) - \psi(\tilde{\gamma}_{0j}^{(1)} + \tilde{\gamma}_{0j}^{(2)})] \\ & + \sum_{k=1}^K q(z_k^{(m)} = 0) \sum_{l=1}^{L_m} q(\xi_{kl}^{(m)} = g) \log f_k(\mathbf{x}_{lk}^{(m)}; \Theta_{k(0,j)}) \end{aligned} \right);$$

$$q(\zeta_{kg}^{(m)} = j) \propto \exp \left(\begin{aligned} & \sum_{h=1}^{j-1} [\psi(\tilde{\gamma}_{kh}^{(2)}) - \psi(\tilde{\gamma}_{kh}^{(1)} + \tilde{\gamma}_{kh}^{(2)})] + [\psi(\tilde{\gamma}_{kj}^{(1)}) - \psi(\tilde{\gamma}_{kj}^{(1)} + \tilde{\gamma}_{kj}^{(2)})] \\ & + \sum_{k=1}^K q(z_k^{(m)} = 1) \sum_{l=1}^{L_m} q(\xi_{kl}^{(m)} = g) \log f_k(\mathbf{x}_{lk}^{(m)}; \Theta_{k(1,j)}) \end{aligned} \right) \quad (29)$$

where $\psi(A) = \frac{\partial}{\partial A} \log \Gamma(A)$ and $\Gamma(\cdot)$ represents the Gamma function.

- Update $\xi_{k'l}^{(m)}$ for $k' = 0, 1, \dots, K$; $l = 1, 2, \dots, L_m$ and $m = 1, 2, \dots, M$ according to a multinomial distribution with probabilities

$$q(\xi_{0l}^{(m)} = g) \propto \exp \left(\begin{aligned} & \sum_{h=1}^{g-1} [\psi(\tilde{\alpha}_{0h}^{(m,2)}) - \psi(\tilde{\alpha}_{0h}^{(m,1)} + \tilde{\alpha}_{0h}^{(m,2)})] \\ & + [\psi(\tilde{\alpha}_{0g}^{(m,1)}) - \psi(\tilde{\alpha}_{0g}^{(m,1)} + \tilde{\alpha}_{0g}^{(m,2)})] \\ & + \sum_{k=1}^K q(z_k^{(m)} = 0) \sum_{j=1}^J q(\zeta_{0g}^{(m)} = j) \log f_k(\mathbf{x}_{lk}^{(m)}; \Theta_{k(0,j)}) \end{aligned} \right);$$

$$q(\xi_{kl}^{(m)} = g) \propto \exp \left(\begin{aligned} & \sum_{h=1}^{g-1} [\psi(\tilde{\alpha}_{kh}^{(m,2)}) - \psi(\tilde{\alpha}_{kh}^{(m,1)} + \tilde{\alpha}_{kh}^{(m,2)})] \\ & + [\psi(\tilde{\alpha}_{kg}^{(m,1)}) - \psi(\tilde{\alpha}_{kg}^{(m,1)} + \tilde{\alpha}_{kg}^{(m,2)})] \\ & + \sum_{k=1}^K q(z_k^{(m)} = 1) \sum_{j=1}^J q(\zeta_{kg}^{(m)} = j) \log f_k(\mathbf{x}_{lk}^{(m)}; \Theta_{k(1,j)}) \end{aligned} \right) \quad (30)$$

- Update $z_k^{(m)}$ for $k = 1, 2, \dots, K$, and $m = 1, 2, \dots, M$ according to a Bernoulli distribution with probabilities

$$\begin{aligned}
q(z_k^{(m)} = 0) &\propto \exp \left(\frac{[\psi(\tilde{\beta}_{k1}) - \psi(\tilde{\beta}_{k1} + \tilde{\beta}_{k2})] + \sum_{l=1}^{L_m} \sum_{g=1}^T}{\sum_{j=1}^J q(\zeta_{0g}^{(m)} = j)q(\xi_{0l}^{(m)} = g) \log f_k(\mathbf{x}_{lk}^{(m)}; \Theta_{k(0,j)})} \right); \\
q(z_k^{(m)} = 1) &\propto \exp \left(\frac{[\psi(\tilde{\beta}_{k2}) - \psi(\tilde{\beta}_{k1} + \tilde{\beta}_{k2})] + \sum_{l=1}^{L_m} \sum_{g=1}^T}{\sum_{j=1}^J q(\zeta_{kg}^{(m)} = j)q(\xi_{kl}^{(m)} = g) \log f_k(\mathbf{x}_{lk}^{(m)}; \Theta_{k(1,j)})} \right) \quad (31)
\end{aligned}$$

The local maximum of the lower bound $\mathcal{L}(q)$ is achieved by iteratively updating the parameters of the variational distributions $q(\cdot)$ according to the above equations. Each iteration guarantees to either increase the lower bound or leave it unchanged. We terminate the algorithm when the change in $\mathcal{L}(q)$ is negligibly small. $\mathcal{L}(q)$ can be computed by substituting the updated $q(\cdot)$ and the prior distributions $p(\Phi|\Psi)$ into (21). To mitigate sensitivity of VB to initialization considerations and local-optimal solutions, we typically run the VB algorithm multiple times with different initializations, and select the result with the maximum lower bound.

In the analysis considered here we perform VB inference on the original model, since this was relatively efficient computationally. One may alternatively perform VB inference on the collapsed model (MacEachern, 1998; Teh et al., 2007), as applied to the combination of collapsed and blocked Gibbs sampling discussed in detail above.

5. Simulation Example

We first consider simple synthesized data to illustrate the HLPP model, with the data set shown in Figure 1. The data are generated with $\mathbf{x}_{lk}^{(m)} \sim \mathcal{N}(\boldsymbol{\mu}_{lk}^{(m)}, \boldsymbol{\Sigma}_{lk}^{(m)-1})$, with $K = 5$, $M = 4$, $L_m = 10$, and

$$\begin{aligned}
\boldsymbol{\mu}_{l1}^{(m)} &= [-2 \quad 2]^T; \quad l = 1, 2, \dots, 5; \quad m = 1, 2; \\
\boldsymbol{\mu}_{l1}^{(m)} &= [2 \quad -2]^T; \quad l = 6, 7, \dots, 10; \quad m = 1, 2; \\
\boldsymbol{\mu}_{l1}^{(m)} &= [2 \quad 2]^T; \quad l = 1, 2, \dots, 5; \quad m = 3, 4; \\
\boldsymbol{\mu}_{l1}^{(m)} &= [-2 \quad -2]^T; \quad l = 6, 7, \dots, 10; \quad m = 3, 4; \\
\boldsymbol{\mu}_{lk}^{(m)} &= [-5 \quad 5]^T; \quad k = 2, 3, 4; \quad l = 1, 2, \dots, 5; \quad m = 1, 2; \\
\boldsymbol{\mu}_{lk}^{(m)} &= [5 \quad -5]^T; \quad k = 2, 3, 4; \quad l = 6, 7, \dots, 10; \quad m = 1, 2; \\
\boldsymbol{\mu}_{lk}^{(m)} &= [5 \quad 5]^T; \quad k = 2, 3, 4; \quad l = 1, 2, \dots, 5; \quad m = 3, 4; \\
\boldsymbol{\mu}_{lk}^{(m)} &= [-5 \quad -5]^T; \quad k = 2, 3, 4; \quad l = 6, 7, \dots, 10; \quad m = 3, 4; \\
\boldsymbol{\mu}_{l5}^{(m)} &= [-10 \quad -10]^T; \quad l = 1, 2, \dots, 10; \quad m = 1, 3; \\
\boldsymbol{\mu}_{l5}^{(m)} &= [10 \quad 10]^T; \quad l = 1, 2, \dots, 10; \quad m = 2, 4; \\
\boldsymbol{\Sigma}_{lk}^{(m)} &= \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}; \quad k = 1, 2, \dots, 5; \quad l = 1, 2, \dots, 10; \quad m = 1, 2, 3, 4.
\end{aligned}$$

These data are constructed with the following motivations. From Figure 1 we note that for each task modalities 1-4 are each characterized by two mixture components, with

modality 5 characterized by a single task-dependent cluster. Note also that modalities 1-4 have the same mixture construction in tasks 1 and 2, as well as (separately) the same mixture construction in tasks 3 and 4. The task-dependent properties of modality 5 are distinct from those of modalities 1-4. Therefore, we anticipate that modalities 1-4 will exhibit “global” clustering, with tasks 1 and 2 clustered together, and similarly tasks 3 and 4 clustered together. By contrast, modality 5 is expected to exhibit distinct “local” clustering, with tasks 1 and 3 clustered together, and tasks 2 and 4 clustered together. Global clustering across all five modalities is clearly inappropriate. However, one may in principle cluster the data independently across tasks for each of the five modalities; this has motivated the construction of the data in modality 1. Note that the differences in the mixture components for modality 1 are far more subtle than those associated with modalities 2-4. Below we demonstrate that when an independent HDP prior is used for task-dependent clustering for each of the modalities, the subtleties associated with modality 1 are missed. By contrast, the HLPP captures the clear clusterings associated with modalities 2-4, and in so doing “pulls along” the same task clustering for modality 1, thereby improving the mixture model for modality 1. The HLPP also allows idiosyncratic “local” clustering for modality 5.

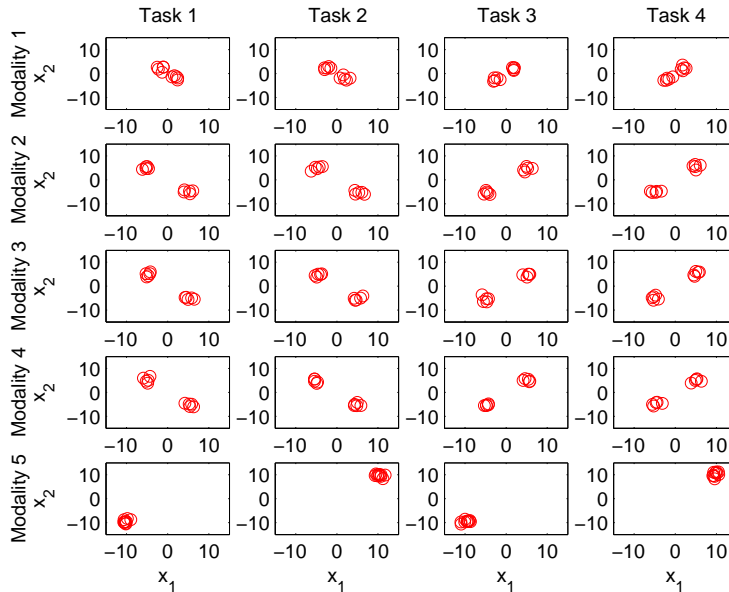


Figure 1: Synthetic data, where x_1 and x_2 represent components of the vector \mathbf{x} .

The HLPP model is implemented for this synthetic data set, with $J = 50$ and $T = 50$. The base distribution is specified as $H_k = \prod_{n=1}^2 [\mathcal{N}(\mu_{kn}; r_{0n}, t_0 \Sigma_{kn}) - \text{Ga}(\Sigma_{kn}; d_0, s_{0n})]$ for $k = 1, 2, \dots, 5$, with $r_{0n} = 0$, $t_0 = 0.01$, $d_0 = 4$, and $s_{0n} = 1$. We place Gamma priors $\text{Ga}(10^{-6}, 10^{-6})$ on α and γ , $\text{Ga}(10^{-6}, 1)$ on β . These parameters were not tuned, and many different settings yield comparable results. The MCMC algorithm described in Section 4.2 is used to obtain samples of posteriors under the HLPP. The results shown below are based on 10,000 collection samples obtained after a burn-in period of 5,000 iterations. Rapid

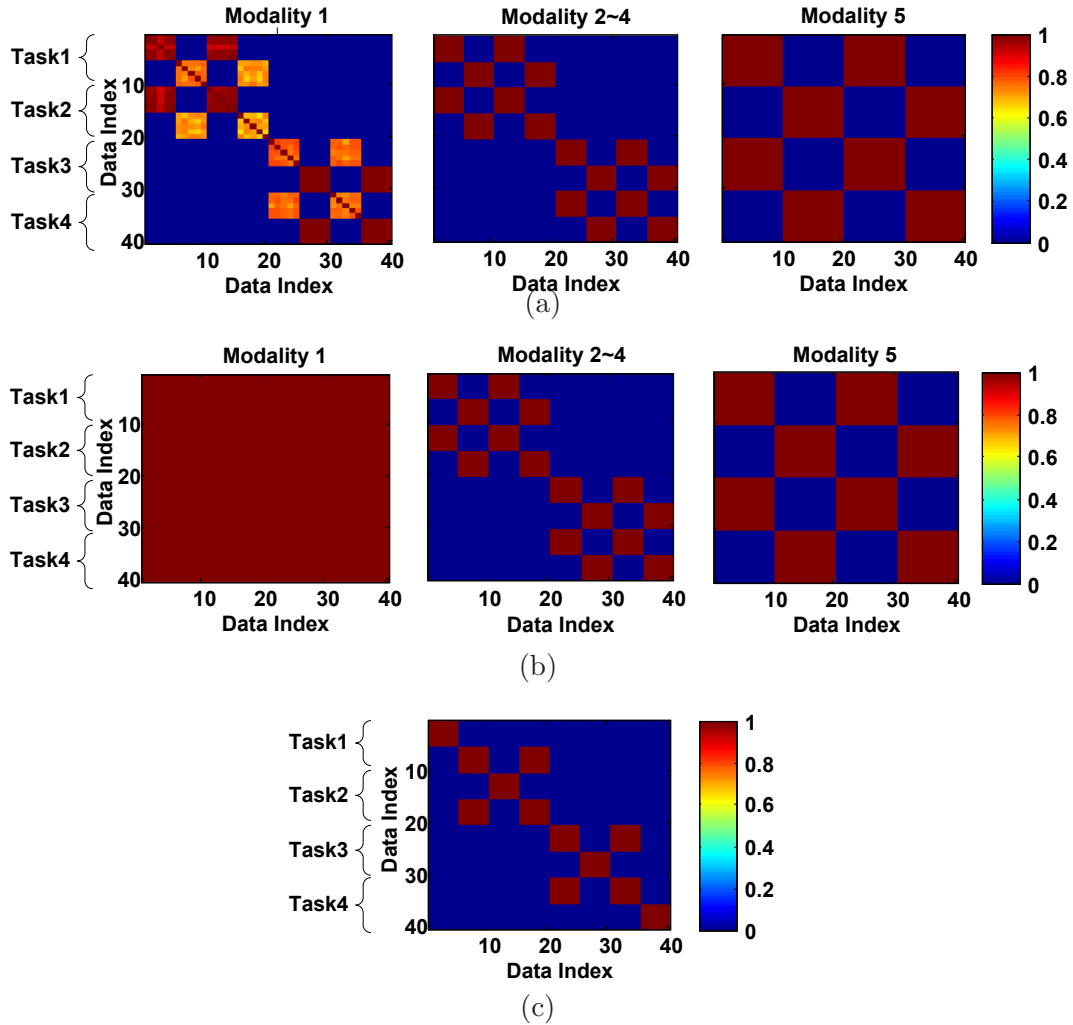


Figure 2: Pairwise posterior probabilities of two data samples being assigned to the same cluster for the simulation example analyzed using the HLPP, independent HDP and HDP models. (a) HLPP results; (b) independent HDP results; (c) “global” HDP results.

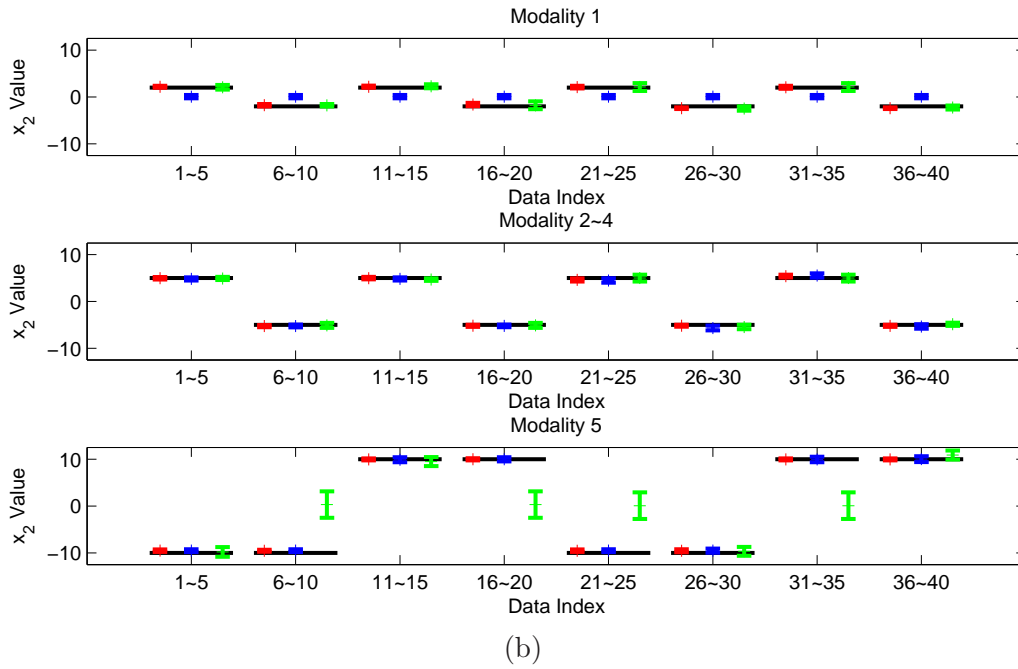
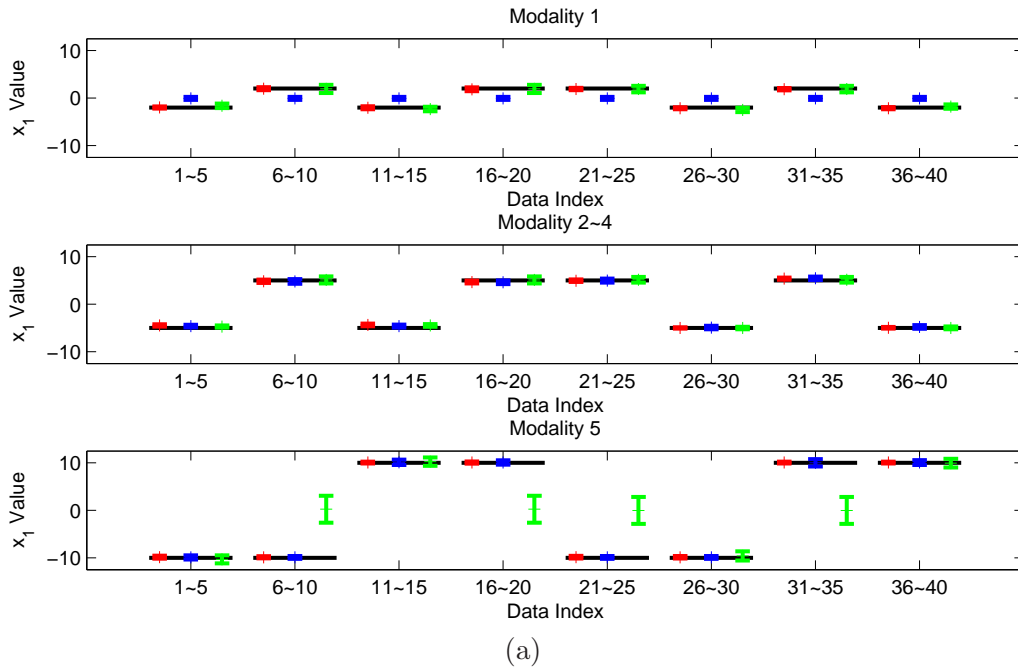


Figure 3: Posterior means and 95% credible intervals for the mean parameters of each modality in the simulation example. The red solid bar indicates the estimates for the HLPP model, the blue shows those for the independent HDP model, the green shows those for the HDP model, and the dark shows the true values of the mean parameters. (a) the first dimension of the data; (b) the second dimension of the data.

convergence has been observed in the diagnostic tests described in (Geweke, 1992) and (Raftery and Lewis, 1992).

From the indicators $\{z_k^{(m)}\}_{k=1,m=1}^{5,4}$ in the HLPP model, we observed that tasks 1-4 selected global clustering and task 5 selected local clustering. Since the two clusters in modality 1 are subtle, the good clustering performance of modalities 2-4 helps modality 1 obtain the correct clustering, as indicated above. Figure 2(a) plots the posterior probability of two data samples being assigned to the same cluster separately for each of the five modalities, as inferred via the HLPP model. It is apparent that the true clustering structure is well represented by the HLPP model. For comparison, Figures 2(b) and 2(c) present the corresponding results obtained by the independent HDP model (special case with $\beta_k \rightarrow \infty$) and global HDP model (special case with $\beta_k \rightarrow 0$), respectively. The similarity matrix of the independent HDP model is not correct for modality 1, impacting model performance as discussed further below. Because of the idiosyncratic clustering structure of modality 5, the similarity matrix of the HDP model is very poor. Figure 3 shows the inferred data-specific posterior means and 95% credible intervals for the mean parameters $\{\mu_{lk}^{(m)}\}_{l=1,k=1,m=1}^{10,5,4}$ via the HLPP model (red), along with the results for the independent HDP model (blue) and the HDP model (green). The black lines represent the true values of the mean parameters. It is clear the posterior densities via the HLPP model are concentrated around the true values, while there are mistakes for modality 1 via the independent HDP model and for modality 5 via the HDP model. In addition, the 95% credible intervals from the HLPP analysis are the tightest, compared with the independent HDP and HDP analysis; this is attributed to proper global and local sharing explicitly imposed by HLPP.

6. Gene-Expression Analysis Application

To further illustrate the proposed HLPP model, we consider an application to gene-expression data, here for a Dengue virus study (Fink et al., 2007). The HLPP is applied to sparse factor analysis (Carvalho et al.; Pournara and Wernish, 2007; Fokoue, 2004; Knowles and Ghahramani, 2007), where the latent factors correspond to the “modalities”. The factor scores represent the data, although now these data are latent. For the Dengue gene-expression data under study we have expression values at multiple time points after cell exposure with the virus, and each of the times represents a task; for each task (time) we have multiple gene-expression samples (from cells exposed to either live or heat inactivated virus). We wish to impose the belief that the factor scores, which can be interpreted as meta-gene expression levels underlying many genes in a pathway, cluster across tasks (time). However, as different meta-genes are involved in different biologic pathways and have varying relationships with viral exposure, the factor scores may vary in their temporal clustering structure. Therefore, we do not expect all components of the factor scores to cluster in the same way. This anticipated phenomenon mitigates the use of global HDP clustering (*i.e.*, the same clustering across all factor scores). However, if we cluster the individual components of the factor scores independently, we do not account for anticipated statistical correlation. This therefore motivates the HLPP construction.

6.1 Sparse Factor Analysis for Gene Expression Data

Assume $\mathbf{y}_l^{(m)}$ is an $N \times 1$ vector representing the gene expressions for N genes, for cell l at time m , with $m = 1, 2, \dots, M$ and $l = 1, 2, \dots, L_m$. It is assumed that $\mathbf{Y} = [\mathbf{y}_1^{(1)} \dots \mathbf{y}_{L_1}^{(1)} \dots \mathbf{y}_1^{(M)} \dots \mathbf{y}_{L_M}^{(M)}]$ is already normalized to zero mean in each row. We consider the case $N \gg K = \sum_{m=1}^M L_m$, which is typical. The parametric sparse factor analysis model (Pournara and Wernish, 2007) characterizes dependence in the high-dimensional gene expression measurements using

$$\begin{aligned}
 \mathbf{y}_l^{(m)} &= \mathbf{B}\boldsymbol{\lambda}_l^{(m)} + \boldsymbol{\epsilon}_l^{(m)}; \quad l = 1, 2, \dots, L_m; \quad m = 1, 2, \dots, M \\
 \boldsymbol{\lambda}_l^{(m)} &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}); \quad l = 1, 2, \dots, L_m; \quad m = 1, 2, \dots, M \\
 B_{nk} &\sim \mathcal{N}(0, \eta_{nk}^{-1}); \quad n = 1, 2, \dots, N; \quad k = 1, 2, \dots, K \\
 \eta_{nk} &\sim \text{Ga}(a_0, b_0); \quad n = 1, 2, \dots, N; \quad k = 1, 2, \dots, K \\
 \epsilon_{nl}^{(m)} &\sim \mathcal{N}(0, \varphi_n^{-1}); \quad l = 1, 2, \dots, L_m; \quad m = 1, 2, \dots, M; \quad n = 1, 2, \dots, N \\
 \varphi_n &\sim \text{Ga}(g_0, h_0); \quad n = 1, 2, \dots, N
 \end{aligned} \tag{32}$$

where B_{nk} is the (n, k) -component of the factor-loading matrix \mathbf{B} , and $\boldsymbol{\lambda}_l^{(m)}$ represents the factor score associated with cell l at time m . Typically one sets K , with $K \ll N$. Note that the components of \mathbf{B} are drawn independently from a Student-t distribution, and hence with appropriate choice of a_0 and b_0 the matrix \mathbf{B} is sparse. Alternatively, one may use a “spike-slab” sparseness construction, as in (Carvalho et al.); (Pournara and Wernish, 2007) gives a comprehensive review of different sparseness priors used for sparse factor analysis. Both \mathbf{B} and all $\boldsymbol{\lambda}_l^{(m)}$ are inferred by the model simultaneously. As discussed in (Carvalho et al.), since \mathbf{B} is “sparse”, which means many of the elements of \mathbf{B} are close to zero, each column ideally will represent a particular biological “pathway”, composed of a relatively small number of relevant genes related to a given latent factor, which correspond to those having factor loadings not close to zero. This is discussed further when presenting results for the time-evolving Dengue virus data (Fink et al., 2007).

6.2 The Extended Factor Model with the HLPP Prior

In the HLPP factor analysis (HLPP-FA) model, the construction of \mathbf{B} and $\boldsymbol{\epsilon}_l^{(m)}$ is unchanged from above, and therefore for simplicity it is not repeated below. What is different is the manner in which the factor scores $\boldsymbol{\lambda}_l^{(m)}$ are drawn from the prior. Specifically, we have

$$\begin{aligned}
 \lambda_{lk}^{(m)} &\sim \mathcal{N}(\mu_{lk}^{(m)}, \Sigma_{lk}^{(m)-1}), \quad (\mu_{lk}^{(m)}, \Sigma_{lk}^{(m)}) \sim \text{LPP}(\boldsymbol{\beta}, G^{(m)}, \{G_k^{(m)}\}_{k=1}^K) \\
 &\quad l = 1, 2, \dots, L_m; \quad k = 1, 2, \dots, K; \quad m = 1, 2, \dots, M \\
 G_k^{(m)} &\sim \text{HDP}(\alpha_k, \gamma_k, H_k); \quad k = 1, 2, \dots, K; \quad m = 1, 2, \dots, M \\
 G^{(m)} &\sim \text{HDP}(\alpha_0, \gamma_0, \prod_{k=1}^K H_k); \quad m = 1, 2, \dots, M \\
 H_k &= \mathcal{N} - \text{Ga}(r_0, t_0, d_0, s_0); \quad k = 1, 2, \dots, K
 \end{aligned} \tag{33}$$

where here $(\mu_{lk}, \Sigma_{lk}) = \theta_{lk}$. Note that (32) assumes that the latent factors are normally distributed and have the same distribution at the different times. In contrast, (33) charac-

terizes the distribution of the latent factors using a flexible mixture of normals, which can vary over time, while borrowing information. This borrowing is accomplished by incorporating the same mean and variance in the normal mixture components at the different times through the HDP structure, while also flexibly clustering latent factors over time through favoring allocation to the same mixture component. However, the LPP structure does not force a latent factor to be allocated to the same component at all times, but allows occasional local switching of components. This is important in flexibly characterizing changes that can occur as a result of the virus.

In the following examples, we also put Gamma priors on α , β and γ . The model in (33) is a direct combination of (5) and (32), and therefore the MCMC inference can also be directly derived based on Section 4.2 and (Fokoue, 2004). Here we only give the main modification concerning sampling latent variable $\lambda_{lk}^{(m)}$.

- Let $\mathbf{y}_{lk}^{(m)*} = \mathbf{y}_l - \sum_{k'' \neq k} \mathbf{B}_{k''} \lambda_{lk''}^{(m)}$ for $k = 1, 2, \dots, K$; $l = 1, 2, \dots, L_m$ and $m = 1, 2, \dots, M$. Since $\mathbf{y}_{lk}^{(m)*} \sim \mathcal{N}(\mathbf{B}_k \lambda_{lk}^{(m)}, \text{diag}^{-1}(\boldsymbol{\varphi}))$, we can sample $\lambda_{kl}^{(m)}$ from

$$\left\{ \begin{array}{l} p(\lambda_{lk}^{(m)} | \dots) \propto \mathcal{N} \left(\begin{array}{l} (\mathbf{B}_k^T \text{diag}(\boldsymbol{\varphi}) \mathbf{B}_k + \Sigma_{k(0, \zeta_{0\xi_{0l}}^{(m)})}^{*-1})^{-1} \\ (\mathbf{B}_k^T \text{diag}(\boldsymbol{\varphi}) \mathbf{y}_{lk}^{(m)*} + \Sigma_{k(0, \zeta_{0\xi_{0l}}^{(m)})}^*) \mu_{k(0, \zeta_{0\xi_{0l}}^{(m)}}^*), \\ (\mathbf{B}_k^T \text{diag}(\boldsymbol{\varphi}) \mathbf{B}_k + \Sigma_{k(0, \zeta_{0\xi_{0l}}^{(m)})}^{*-1})^{-1} \end{array} \right) \text{ if } z_k^{(m)} = 0 \\ p(\lambda_{lk}^{(m)} | \dots) \propto \mathcal{N} \left(\begin{array}{l} (\mathbf{B}_k^T \text{diag}(\boldsymbol{\varphi}) \mathbf{B}_k + \Sigma_{k(1, \zeta_{k\xi_{kl}}^{(m)})}^{*-1})^{-1} \\ (\mathbf{B}_k^T \text{diag}(\boldsymbol{\varphi}) \mathbf{y}_{lk}^{(m)*} + \Sigma_{k(1, \zeta_{k\xi_{kl}}^{(m)})}^*) \mu_{k(1, \zeta_{k\xi_{kl}}^{(m)}}^*), \\ (\mathbf{B}_k^T \text{diag}(\boldsymbol{\varphi}) \mathbf{B}_k + \Sigma_{k(1, \zeta_{k\xi_{kl}}^{(m)})}^{*-1})^{-1} \end{array} \right) \text{ if } z_k^{(m)} = 1 \end{array} \right. \quad (34)$$

The symbols correspond to those used in (10), where $(\mu_{k(i,j)}^*, \Sigma_{k(i,j)}^*) = \Theta_{k(i,j)}$. The remaining variables specified in (33) are sampled in a similar manner as in Section 4.2 and (Fokoue, 2004).

6.3 Experimental Results for Gene-Expression Data

The Dengue data considered here is time-evolving expression data, and is publicly available at <http://www.ncbi.nlm.nih.gov/projects/geo> (accession number is GSE6048). The data consist of six groups (tasks) of samples measured at six time points. The six time points are 3, 6, 12, 24, 48 and 72 hours after a HepG2 cell is exposed to the NGC Dengue virus (both live and heat-inactivated/dead viruses are used). At each time point the transcriptome is profiled using biological repeats. The corresponding number of samples at the six time points are 10, 12, 12, 12, 12, 11. The number of genes in this dataset is 20,160. The detailed description of these genes can be found in (Fink et al., 2007). In the plots that follow, the samples are ordered according to their time points, from early to late time. One question of interest is whether cells with live virus have a systematically different profile

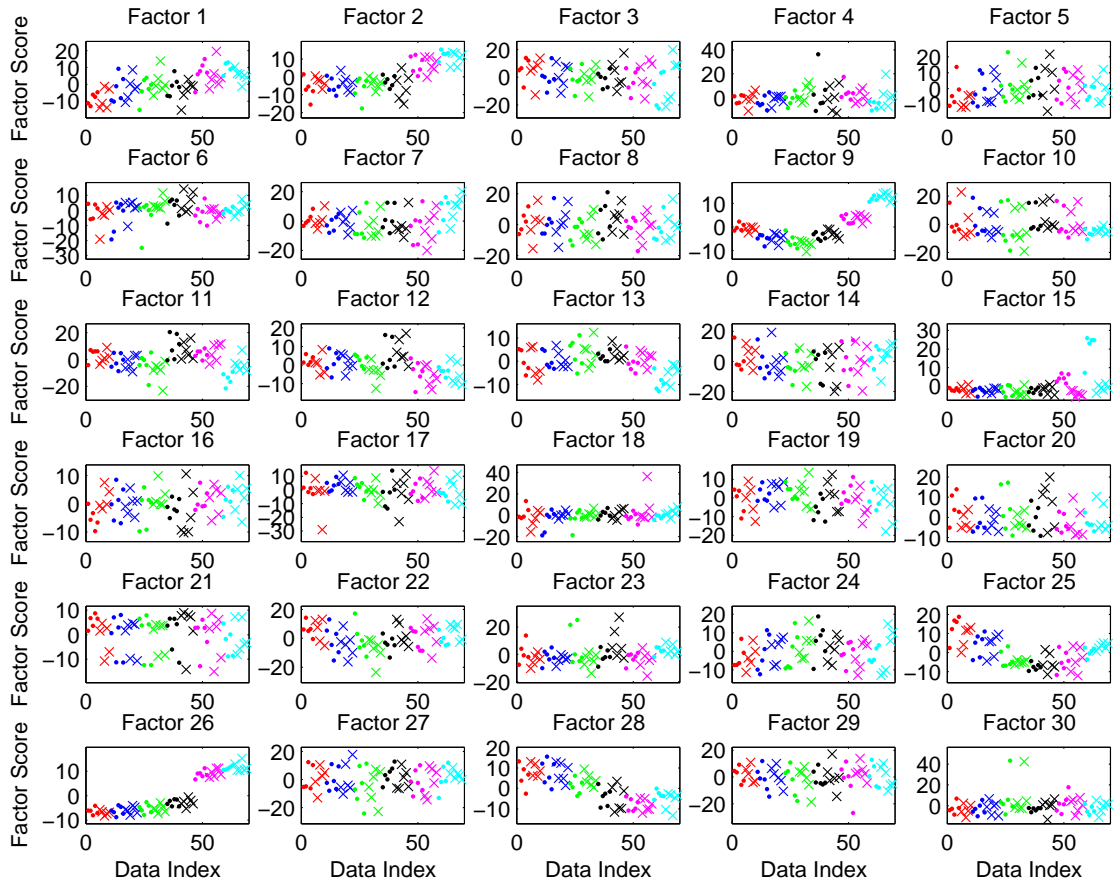


Figure 4: Posterior means for the components of each factor score, as computed via the sparse factor analysis model in (Carvalho et al.). Different color denote different time points (tasks), time increases from left to right across the horizontal axis; “.” represents the cell cultures exposed to live virus, and “x” represents the cell cultures exposed to heat inactivated virus.

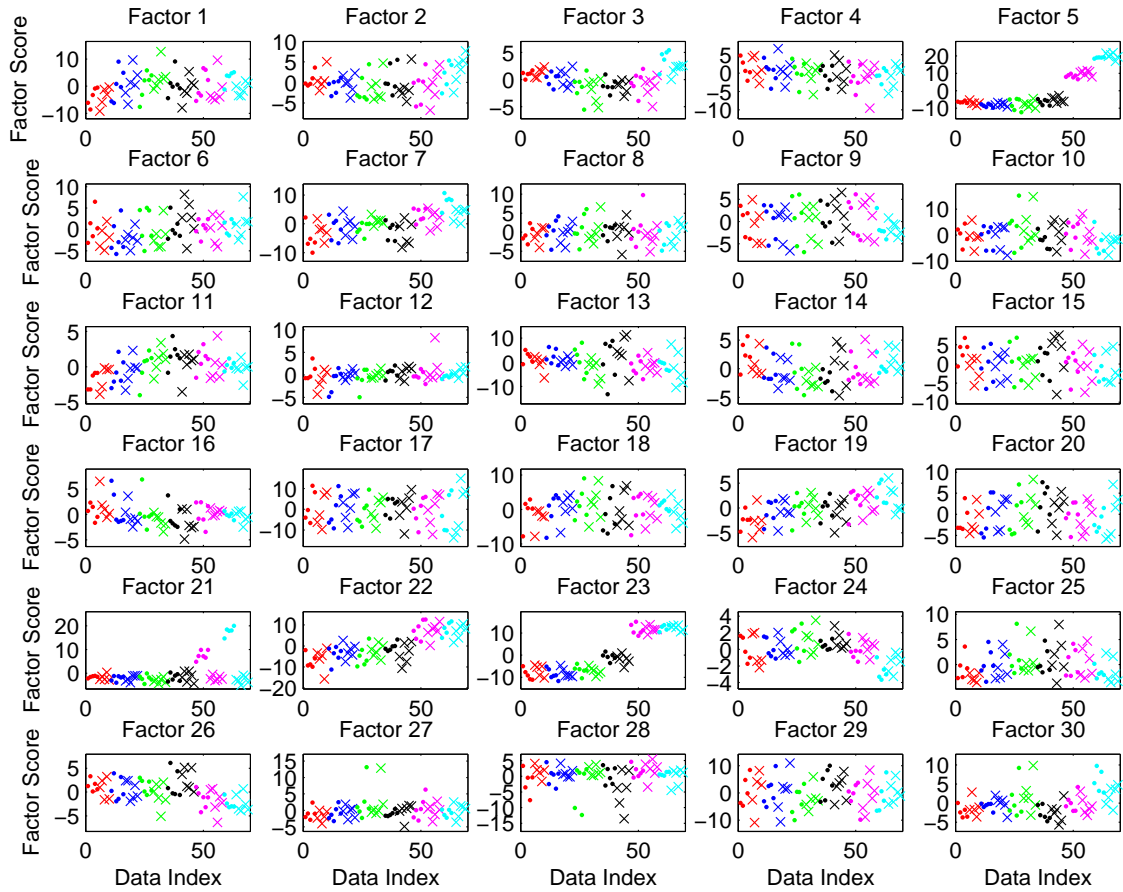


Figure 5: Posterior means for the components of each factor score, as computed via the sparse factor analysis model with the HLPP prior, where different color denote different time points (tasks), time increases from left to right across the horizontal axis; “.” represents the cell cultures exposed to live virus, and “x” represents the cell cultures exposed to heat inactivated virus.

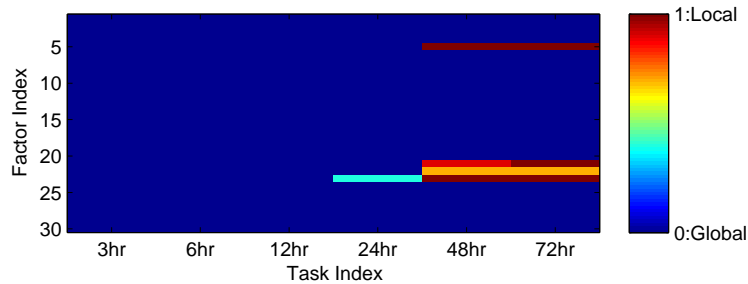


Figure 6: The average indicator matrix for selecting global or local clustering via the sparse factor analysis model with the HLPP prior.

of changes in the latent factor score distributions over time relative to cells with dead virus. For factors exhibiting such changes, it is also of interest to identify the associated genes (*i.e.*, those having loadings not close to zero) and assess whether they have known biological significance. In (Fink et al., 2007), differential gene expression was noted between cells exposed to live or heat inactivated virus at 48 hours post exposure onset.

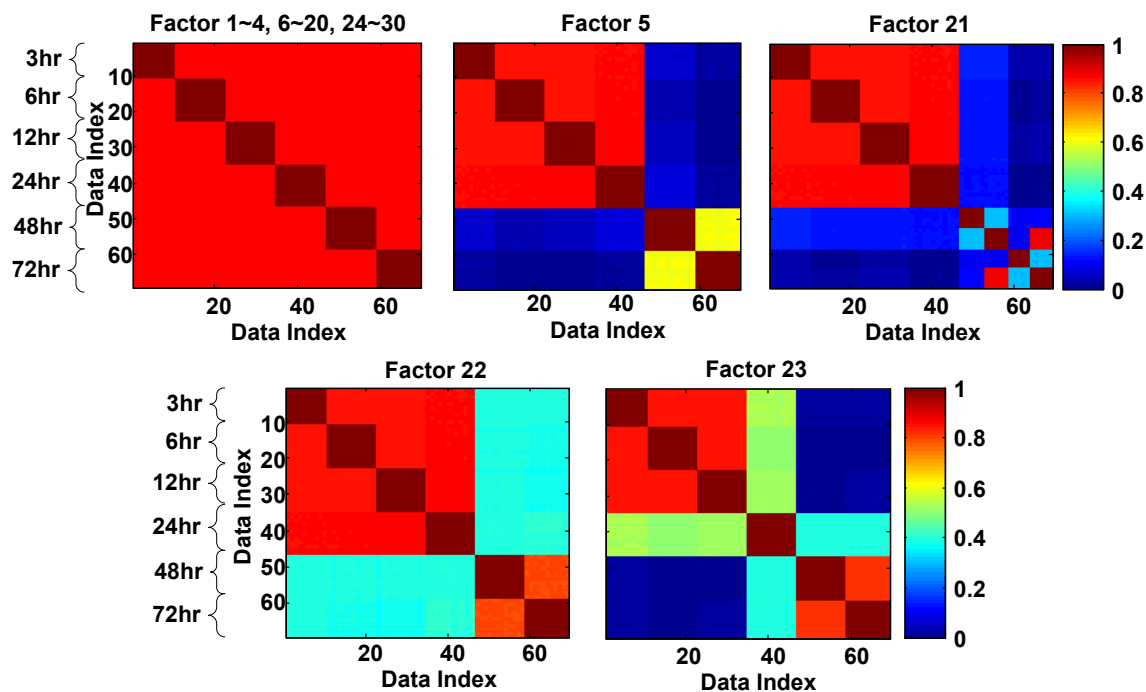


Figure 7: Pairwise posterior probabilities of two data samples being assigned to the same cluster for the Dengue data, analyzed using the sparse factor analysis model with the HLPP prior.

The traditional sparse factor analysis model (FA) in (32) and the sparse factor analysis model with the HLPP prior (HLPP-FA) are applied to the normalized gene expression data using $K = 30$ factors, with truncation levels $J = 50$ and $T = 50$; in (32) all factors $\lambda_t^{(m)}$ are drawn independently. The hyper-parameters are $r_{0k} = 0$ for $k = 1, 2, \dots, 20$, $t_0 = 0.01$, $d_0 = 4$, $s_{0k} = 1$ for $k = 1, 2, \dots, 20$, $a_0 = 0.1$, $b_0 = 10^{-6}$, $g_0 = 10^{-6}$ and $h_0 = 10^{-6}$. We also place Gamma priors $\text{Ga}(10^{-6}, 10^{-6})$ on α and γ , and $\text{Ga}(10^{-6}, 1)$ on β . These hyper-parameters were not optimized, and the results are relatively insensitive to most “reasonable” settings. The results given below are based on 10,000 samples collected from the combination of collapsed and blocked Gibbs sampling, after a burn-in period of 5,000 iterations. Rapid convergence has been observed in the diagnostic tests as described in (Geweke, 1992) and (Raftery and Lewis, 1992).

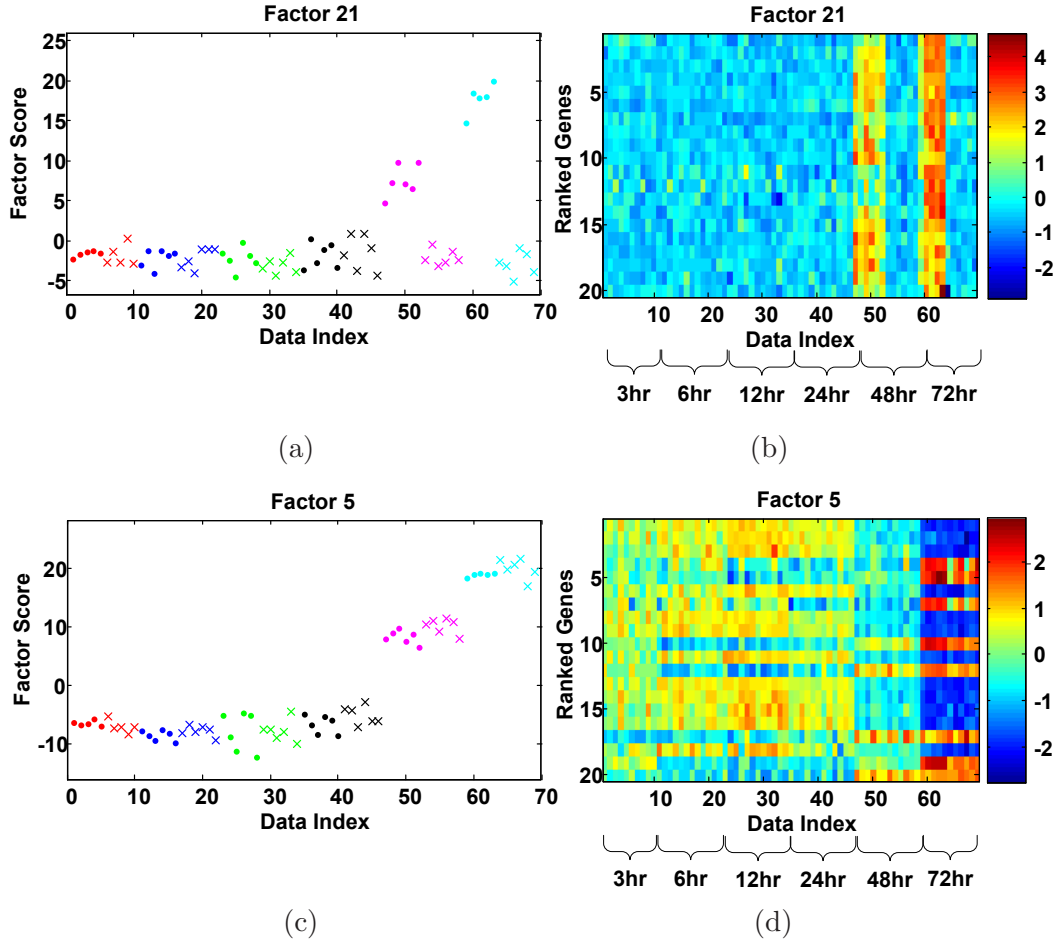


Figure 8: Two important factors and corresponding top-20 genes via the sparse factor analysis model with the HLPP prior. These factors are deemed important because they are characterized by factor scores that exhibit local/distinctive clustering with increasing time. (a) posterior means for Factor 21, where different colors denote different time points, time increases from left to right across the horizontal axis; “.” represents the cell cultures exposed to live virus, and “x” represents the cell cultures exposed to heat inactivated virus; (b) top 20 genes contributing to Factor 21; (c) posterior means for Factor 5, where different color denote different time points, time increases from left to right across the horizontal axis; “.” represents the cell cultures exposed to live virus, and “x” represents the cell cultures exposed to heat inactivated virus; (d) top-20 genes contributing to Factor 5.

Table 1: Detailed gene description of the top 20 genes corresponding to Factor 21 via HLPP-FA. These genes are highly correlated with viral infection (per analysis by the fifth author), and are found to be differentially expressed between the experimental sets in (Fink et al., 2007).

Name	Group	Description
IL8	NFkB related	Chemokine activity, attracts neutrophils, basophils, and t-cells
MDA5	Interferon related	interferon induced with helicase C domain 1;IFIH1
HERC5	Ubiquitin related	Ubiquitin-protein ligase, cyclin-E binding protein 1
G1P2	Interferon related	Proteolysis, interferon-stimulated protein, 15 kDa (ISG15)
IFIT3-1	Interferon related	Interferon-induced protein with tetratricopeptide repeats 3
IFIT1	Interferon related	interferon-induced protein with tetratricopeptide repeats 1;IFIT1
ATF3	Interferon related	mRNA transcription regulation;Induction of apoptosis
VIPERIN-1	Interferon related	Virus inhibitory, endoplasmic reticulum-associated, interferon inducible
IFNB1	Interferon related	interferon, beta 1, fibroblast
MX1	Interferon related	myxovirus (influenza virus) resistance 1, interferon-inducible protein p78
KRT17		Intermediate filament;Structural protein
LOC93082		ortholog of mouse lung-inducible C3HC4 RING domain protein
IFIT2	Interferon related	Interferon-induced protein with tetratricopeptide repeats 2
IP10	NFkB related	Cytokine and chemokine signaling;Macrophage-mediated immunity
LGP2		Nucleoside, nucleotide and nucleic acid metabolism
OAS1	Interferon related	nucleotide and nucleic acid metabolism;Interferon-mediated immunity
VIPERIN-2	Interferon related	Virus inhibitory, endoplasmic reticulum-associated, interferon inducible
OAS2	Interferon related	nucleotide and nucleic acid metabolism;Interferon-mediated immunity
CCL5	NFkB related	Cytokine and chemokine mediated signaling pathway
I-TAC	NFkB related	Cytokine and chemokine signaling;Macrophage-mediated immunity

The posterior means of the factor scores, computed for all 30 factors via the two methods, are shown in detail in Figures 4 and 5 (different colors denote different time points, time increases from left to right across the horizontal axis; “.” represents the cell cultures exposed to live virus, and “x” represents the cell cultures exposed to heat inactivated virus), from which we observe that factor 15 via FA and factor 21 via HLPP-FA distinguish between cells exposed to live virus and cells exposed to heat-inactivated virus, with separation between the groups noted at 48 hours and continuing through 72 hours post inoculation. In addition, three factors (factor 9, factor 26 and factor 28) via FA and (factor 5, factor 22 and factor 23) via HLPP-FA change in a correlated manner with time, and seem to be associated with cell splitting and growth (based upon the associated important genes in the associated factor loadings). All biological interpretation of these results were performed by the fifth author.

Comparing factor 15 inferred via FA and factor 21 inferred via HLPP-FA, it is clear that HLPP-FA enhances the separation of the factor scores for cells exposed to live virus relative to cells exposed to heat-inactivated virus. This is likely because the factor scores in the FA model are drawn from independent normals, while those in the HLPP-FA model are drawn from a flexible mixture of normals. The factor scores associated with factor 21 from HLPP-FA are represented in terms of two clusters at later times, with one cluster

associated with cells with live Dengue virus, and the other cluster associated with heat-inactivated virus. Note that this association was inferred from the model, since knowledge of the different states of the virus was not imposed in the model.

Figure 6 shows the average indicator matrix for selecting global and local clustering via HLPP-FA, where factors 5, 21, 22 and 23 select local clustering after 24 hours, while other factors select global clustering. It is deemed therefore that factors 5, 21, 22 and 23 are associated with the idiosyncratic (time-evolving) properties of the virus as well as non-virus-related cell splitting and growth (perhaps less evident in the expression data before 24 hours). The other factors, that do not evolve with time, are by contrast deemed to be associated with cell activities (or other aspects) unrelated or weakly related to the virus nor time evolving properties of the cells (it is important to note that this product is unique to the HLPP-FA, relative to the model in (Carvalho et al.), providing an important tool for interpreting the factors). While one may also infer Dengue-related factors from the FA results shown in Figure 4, one must “eyeball” all the factor scores as a function of time, and the human must decide which ones are of interest. By contrast, HLPP can automatically point to the important factors via the local sharing depicted in Figure 6.

Figure 7 plots the posterior probability of two data samples being assigned to the same cluster for factors 5, 21, 22, 23 and the other factors as computed via the HLPP-FA model. It is apparent that the clustering structure we find in Figure 5 is well represented by the similarity matrices obtained by the posterior indicators in HLPP-FA. In addition, note there are two clusters at 48 and 72 hours for factor 21 (as mentioned above, one associated with live virus, and the other with heat-inactivated virus). This means the data from each of the two tasks are drawn from a mixture distribution rather than a single distribution, which is the main motivation of the HLPP model (as compared to using the original LPP alone).

The posterior means for factor 21 and factor 5 via HLPP-FA are shown more clearly in Figure 8(a) and 8(c). By ranking the absolute values of the corresponding factor loading weights, we infer the top 20 genes contributing to factor 5 and factor 21, with these depicted in Figure 8(b) and 8(d). Table 1 lists the detailed gene description of the top 20 genes in Figure 8(b). These genes are highly correlated with viral infection, and are found to be differentially expressed between the experimental sets in (Fink et al., 2007). We now consider a more-detailed analysis on the selected genes shown in Figures 8(b) and 8(d). Many bioinformatics applications can be used to classify genes represented into categories. One such program, GATHER (Gene Annotation Tool to Help Explain Relationships) groups genes by Gene Ontology categories and provides an assessment of how strongly a GO annotation is associated with a particular gene list (Chang and Nevins, 2006). Notably, factor 21 genes cluster in the GO categories of immune response ($p < 0.0001$), defense response ($p < 0.0001$), and response to biotic stimulus ($p < 0.0001$), as well as response to virus ($p < 0.0001$) and regulation of viral life cycle ($p < 0.0001$). The factor score associated with factor 5 (similar to factors 22 and 23) evolves in a coherent manner post inoculation, but it does not discriminate between cells infected with live versus heat-killed virus. Genes represented in these factors were highly associated with mitosis, the cell cycle and nuclear division ($p < 0.0001$), as would be expected as representative of cellular division in cell culture across time.

For this Dengue data, the combination of collapsed and blocked Gibbs sampling as applied to HLPP-FA and FA required about 1 and 3 hours, respectively, using non-optimized

MatlabTM software on a Pentium IV PC with a 2.1 GHz CPU. We found that VB-based analysis typically yields similar results, but generally MCMC results are more reliable (as a consequence of the local-optimal nature of the VB solution).

7. Image-Annotation Analysis Application

In the above two examples, since the number of tasks was relatively small, we considered MCMC inference. To demonstrate the performance of the VB-HLPP formulation, we consider an application with a large number of annotated images. The assumption in this example is that we are given image-text pairs (annotated images), and we seek to cluster/sort them. Since there are only two modalities (text and image features), and because the number of samples of each of these modalities may be different for an annotated image, we adapt the general HLPP model.

7.1 Modified HLPP Model for Image-Annotation Application

In this setting one “modality” corresponds to the words in an annotation/caption, and the other modality corresponds to features extracted from the image. The model in (5) assumes that the number of observations associated with the different modalities are the same, and moreover that the individual samples characteristic of the different modalities are always observed jointly. We modify this model, such that it is appropriate for the image-text example of interest here. Specifically, let $L_1^{(m)}$ represent the number of observations for modality 1 in task m , with $L_2^{(m)}$ similarly defined for modality 2. The model structure for $K = 2$ and for $L_1^{(m)} \neq L_2^{(m)}$ is

$$\begin{aligned}
 & \begin{cases} \boldsymbol{\theta}_{lk}^{(m)} = \boldsymbol{\vartheta}_{lk}^{(m)} & \text{if } z_k^{(m)} = 0 \\ \boldsymbol{\theta}_{lk}^{(m)} \sim G_k^{(m)} & \text{if } z_k^{(m)} = 1 \end{cases} ; \quad l = 1, \dots, L_k^{(m)}; \quad k = 1, 2; \quad m = 1, \dots, M \\
 & z_k^{(m)} \sim \rho_k \delta_0 + (1 - \rho_k) \delta_1; \quad k = 1, 2, ; \quad m = 1, 2, \dots, M \\
 & \rho_k \sim \text{Beta}(1, \beta_k); \quad k = 1, 2 \\
 & \boldsymbol{\vartheta}_{lk}^{(m)} \sim G^{(m)}|_k; \quad l = 1, \dots, L_k^{(m)}; \quad k = 1, 2; \quad m = 1, \dots, M
 \end{aligned} \tag{35}$$

The expression $\boldsymbol{\vartheta}_{lk}^{(m)} \sim G^{(m)}|_k$ implies that the k th mixture of parameters is drawn from the k th marginal of $G^{(m)}$. The construction of $G^{(m)}$ and $G_k^{(m)}$ is same as that in (5) with $K = 2$, and therefore for simplicity it is not repeated in (35). The model implies that when $z_k^{(m)} = 0$ the *mixture weights* of atoms used for the two modalities are linked, but not the specific atoms associated with a given sample (since the samples from the different modalities are not explicitly linked, for the number of image and text features are not the same). The hyperparameters are set such that it is probable that $z_k^{(m)} = 0$. We note that the special case $z_k^{(m)} = 0$ for all k and m corresponds to the assumptions used in (Barnard et al., 2003; Blei and Jordan, 2003). To distinguish from the HLPP model discussed in Section 3, this modified model is referred to as HLPP*. The MCMC and VB inference formulations for (35) are very similar to those given in Section 4, and are therefore omitted here.

7.2 Feature Extraction for Image-Annotation Data

Let $\{\mathbf{x}_{l_1}^{(m)}\}_{l=1}^{L_1^{(m)}}$ denote the image feature vectors in the m th task, and let $\{\mathbf{x}_{l_2}^{(m)}\}_{l=1}^{L_2^{(m)}}$ denote the corresponding text feature vectors. The image sizes and the number of objects in these images may be different; therefore, the number of feature vectors $L_k^{(m)}$ may vary between tasks and modalities.

For imagery, we employ features constituted by the independent feature subspace analysis (ISA) technique (Hyvärinen and Hoyer, 2000). These features have proven to be relatively shift or translation invariant. In brief, the ISA feature extraction process is composed of two steps: i) We employ patches of images as training data, to estimate several independent feature subspaces via a modification of the independent component analysis (ICA) (Common, 1994). The n th feature subspace with $n = 1, 2, \dots, N_1$ is represented as a set of orthogonal basis vectors, *i.e.*, \mathbf{w}_{nf} with $f = 1, 2, \dots, C$, where C is the dimension of the subspace. ii) The feature \mathbf{F} of a new patch \mathbf{R} is computed as the norm of the projections on the feature subspaces

$$F_n(\mathbf{R}) = \sum_{f=1}^C \langle \mathbf{w}_{nf}, \mathbf{R} \rangle^2; \quad n = 1, 2, \dots, N_1 \quad (36)$$

where N_1 is the number of independent feature subspaces and $\langle \cdot \rangle$ is the inner product. For a patch of arbitrary size, the extracted feature vector is N_1 -dimensional. Interesting invariance properties (Hyvärinen and Hoyer, 2000) enable the ISA features to be widely applied in image analysis. (Hoyer and Hyvärinen, 2000) also discussed how to extract the ISA features from color images. For the extracted ISA feature vectors $\{\mathbf{x}_{l_1}^{(m)}\}_{l=1}^{L_1^{(m)}}$, we assume that each feature vector is drawn from a Gaussian distribution with diagonal covariance, *i.e.*, $\mathbf{x}_{l_1}^{(m)} \sim \mathcal{N}(\boldsymbol{\theta}_{l_1}^{(m)} = (\boldsymbol{\mu}_{l_1}^{(m)}, \boldsymbol{\Sigma}_{l_1}^{(m)}))$ with $\boldsymbol{\mu}_{l_1}^{(m)}$ representing the mean vector and $\boldsymbol{\Sigma}_{l_1}^{(m)}$ representing the diagonal precision matrix, and the base distribution corresponds to the product of a normal-gamma distributions, *i.e.* $H_1 = \prod_{n=1}^{N_1} \left[\mathcal{N}(\boldsymbol{\mu}_{l_1 n}^{(m)}; r_{0n}, t_0 \boldsymbol{\Sigma}_{l_1 n}^{(m)}) - \text{Ga}(\boldsymbol{\Sigma}_{l_1 n}^{(m)}; d_0, s_{0n}) \right]$.

The text is modeled in a manner related to (Blei et al., 2003) and (Blei and Jordan, 2003); specifically, we let $x_{l_2}^{(m)} \in \{1, 2, \dots, N_2\}$ correspond to an index of the l th word in the caption for image m , with N_2 denoting the number of unique words in all the tasks. We let $x_{l_2}^{(m)} \sim \text{Mult}(\boldsymbol{\theta}_{l_2}^{(m)} = \boldsymbol{\varrho}_{l_2}^{(m)})$, with $\boldsymbol{\varrho}_{l_2}^{(m)}$ corresponding to an $N_2 \times 1$ probability vector, *i.e.* a probability mass function (PMF) parameter. As a conjugate choice, we choose the base distribution in the annotation model to correspond to a Dirichlet distribution, *i.e.* $H_2 = \text{Dir}(\boldsymbol{\varrho}_{l_2}^{(m)}; \boldsymbol{\nu}_0)$.

7.3 Similarity Measure

For the image-annotation example considered below, we seek to cluster/sort annotated images, including information from the image features as well as the text. There are many ways this may be performed, while here we count the relative number of times a given image-text pair is characterized by specific atoms (from the stick-breaking representation of the above model). Given a VB run, the posterior atom weights for modality $k \in \{1, 2\}$

and task $m \in \{1, 2, \dots, M\}$ inferred via HLPP* are defined as

$$\begin{aligned} W_{k,j}^{(m)} &= \frac{q(z^{(m)}=0)}{L_k^{(m)}} \sum_{l=1}^{L_k^{(m)}} q(c_{0kl} = j) \\ W_{k,j+J}^{(m)} &= \frac{q(z^{(m)}=1)}{L_k^{(m)}} \sum_{l=1}^{L_k^{(m)}} q(c_{1kl} = j) \end{aligned} \quad (37)$$

where c_{0kl} identifies the atom index for sample l of modality k if $z^{(m)} = 0$, and c_{1kl} is the same when $z^{(m)} = 1$; the function $q(\cdot)$ denotes the posterior probability of the argument. It is assumed that the stick-breaking representation is truncated such that there are J atoms of each type, and hence $\mathbf{W}_k^{(m)}$ corresponds to a $2J \times 1$ probability vector. To measure the similarity of tasks m_1 and m_2 for modality k , we define a kernel function

$$\mathcal{SIM}_{VB}(m_1, m_2 | k) = \exp \left(-\frac{\mathcal{D}^2(\mathbf{W}_k^{(m_1)} \| \mathbf{W}_k^{(m_2)})}{\sigma^2} \right) \quad (38)$$

where σ is a fixed parameter (note that the choice of σ does not change the order of similarities); $\mathcal{D}(\cdot \| \cdot)$ denotes the Kullback-Leibler (KL) distance measure. Since the KL distance is asymmetric, when performing computations this distance is averaged as $[\mathcal{D}_{KL}(\mathbf{W}_k^{(m_1)} \| \mathbf{W}_k^{(m_2)}) + \mathcal{D}_{KL}(\mathbf{W}_k^{(m_2)} \| \mathbf{W}_k^{(m_1)})]/2$, yielding a symmetric distance measure. Furthermore, although the atoms have different meanings for different modalities (*i.e.*, the base measures associated with each modality are different), we may use the combined posterior atom-weight vector over different modalities (after normalization) to measure the similarity between tasks. The detailed approach is to substitute $\mathbf{W}_k^{(m)}$ in (38) with $\mathbf{W}^{(m)} = \frac{1}{2}[\mathbf{W}_1^{(m)\top}, \mathbf{W}_2^{(m)\top}]^\top$ for $m = m_1, m_2$ to get the combined similarity measure $\mathcal{SIM}_{VB}(m_1, m_2)$ (to avoid repeated notation, the explicit expressions for the combined similarity measure are omitted here).

For a new task $\{\mathbf{x}_{lk}^*\}_{l=1, k=1}^{L_k, 2}$, we can predict its posterior weight on each atom associated with each VB run as

$$W_{k,j+i \cdot J} \propto \begin{aligned} &(1-i)\mathbb{E}[\rho] \sum_{l=1}^{L_k} \mathbb{E}[\omega_{0j}] f_k(\mathbf{x}_{lk}^*; \mathbb{E}[\Theta_{k(0,j)}]) \\ &+ i(1-\mathbb{E}[\rho]) \sum_{l=1}^{L_k} \mathbb{E}[\omega_{kj}] f_k(\mathbf{x}_{lk}^*; \mathbb{E}[\Theta_{k(1,j)}]) \end{aligned} \quad (39)$$

where $i = 0, 1$; $j = 1, 2, \dots, J$; $f_k(\cdot)$ is the corresponding parametric model defined in Section 7.2; and $\mathbb{E}[A]$ represents the posterior expectation of the variable A in the given VB run. Then we can measure this task's similarity with other tasks by (38) or its extended form for the combined similarity measure. Thus once we have learned the HLPP* model using one large-scale dataset (essentially learning the set of atoms), we can directly do clustering for new datasets by the predicted posterior atom-weight vectors. Although the learning procedure may be somewhat time-consuming, the *testing* on new data in the sense discussed above is fast.

Before proceeding, we seek to provide insight into why HLPP* may provide better sorting performance for image-annotation pairs than always coupling the text-image atoms (special case of HLPP* when $\rho_k = 0$) or always treating them independently ($\rho_k = 1$). The distance measure between any two image-text pairs, as summarized above, basically reduces to a count of model atoms, and two image-text pairs are deemed similar if the associated atom counts are similar. In the clustering of tasks, the model infers an appropriate set of

atoms representative of the image features and word counts, with this selection of atoms playing a critical role in subsequent sorting (since these atoms drive the subsequent atom counts). If the mixture weights of image and text atoms are always coupled, then there can be difficulties when the (imperfect) image features are similar for two image parts that are consistently characterized by different words. However, there are situations in which the coupling of text features and words adds value, particularly for image features that are only different in subtle ways (recall modality 1 in the synthesized problem considered above). If the image features are different in a subtle way, and the words are consistently similar, then HLPP* encourages the sharing of the subtle image features in a manner consistent with the words. The HLPP* model therefore accounts for subtle but important differences in the image features, which are disambiguated by the word features, while also allowing the image and word models to be treated independently on occasion, when the data deems appropriate.

7.4 Experimental Results

We tested the HLPP* model on a subset of an image-annotation database named “LabelMe”, available at <http://labelme.csail.mit.edu/>. We emphasize that our assumption is that we have access to annotated images, and our objective is to cluster or find inter-relationships between these; we are *not* seeking to automatically annotate images. We select six types of data in this database: highway, building, coast, forest, bedroom and office. Forty tasks are randomly selected from the database for each type, where half the tasks are used to learn the models and the rest are used to examine model prediction performance, yielding a total of 240 tasks. Since the original images had much variability in size, we down-sampled some large images, yielding images of size varying from 256×256 to 300×400 pixels. To capture textural information within the image features, we first divided each image into contiguous 32×32 -pixel non-overlapping patches (more than 20,000 patches in total) and then extracted ISA features from each patch.

In the following we apply both the MCMC and VB algorithms for HLPP* modeling of multiple image-annotation data. The hyper-parameters for the examples that follow are: $\text{Ga}(10^{-6}, 10^{-6})$ for α and γ ; $\text{Ga}(10^{-6}, 1)$ for β ; $r_{0n_1} = 0$, $t_0 = 0.01$, $d_0 = N_1 + 2$, $s_{0n_1} = 1$, $\varrho_{0n_2} = 1/N_2$ for $n_1 = 1, \dots, N_1$ and $n_2 = 1, \dots, N_2$, where $N_1 = 40$ denotes the dimensionality of image-feature, and $N_2 = 147$ denotes the number of unique words with the stopwords removed according to a common stopwords list provided by the dataset website. The DP truncation levels are set to $J = 50$ and $T = 50$. The MCMC results are based on 10,000 samples obtained after a burn-in period of 5,000 iterations. Convergence has been observed in the diagnostics tests described in (Geweke, 1992; Raftery and Lewis, 1992). For VB, the algorithm was run five times and the similarity-measure results with the largest lower bound are selected across these runs, as described in Section 7.1.

We compute the similarity of the posterior atom-weight vectors between any two tasks, as discussed in Section 7.1. Figures 9(a) and 9(b) present the respective MCMC similarity matrices on the posterior atom-weight vectors over image features and annotation features *alone*, while Figures 9(c) and 9(d) show the MCMC and VB similarity matrices on the posterior atom-weight vectors over both image *and* annotation features. As shown in Figures 9(a)-9(d), the combined clustering (image and text) outperforms the clustering over separate

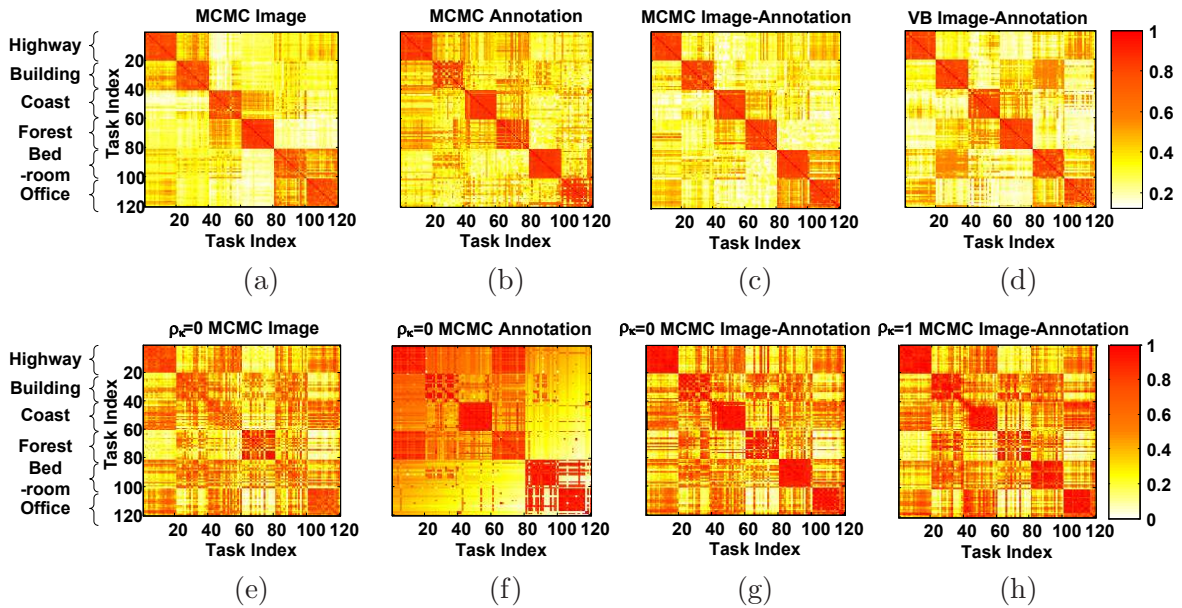


Figure 9: Similarity matrices based on the KL divergences on the posterior atom-weight vectors across tasks via HLPP*. (a) MCMC and using image-feature atoms *alone*; (b) MCMC and using annotation-feature atoms *alone*; (c) MCMC and using both image- *and* annotation-feature atoms; (d) VB and using both image- *and* annotation-feature atoms; (e) $\rho_k = 0$ (*i.e.* independent HDP), MCMC and using *only* image features; (f) $\rho_k = 0$ (*i.e.* independent HDP), MCMC and using *only* annotation features; (g) $\rho_k = 0$ (*i.e.* independent HDP), MCMC and using both image- *and* annotation-feature atoms; (h) $\rho_k = 1$, MCMC and using image-annotation feature atoms (Barnard et al., 2003; Blei and Jordan, 2003); where $k = 1, 2$.

image-feature atoms or annotation-feature atoms. For comparison, Figures 9(e)-9(h) present the corresponding similarity matrices across tasks generated using two extremes of HLPP*: i) $\rho_k = 0$, and ii) $\rho_k = 1$, for $k = 1, 2$. We note that case i) corresponds to treating the image and text features independently in the prior; while case ii) corresponds to the situation in which the *mixture weights* of image and text atoms are always the same, as in (Barnard et al., 2003; Blei and Jordan, 2003). Figures 9(e) and 9(f) only use image features and text features, respectively; Figure 9(g) uses both of the posterior atom-weight vectors via the models of Figures 9(e) and 9(f) to yield the combined similarity matrix (as discussed in Section 7.3).

To provide more-quantitative comparisons, Figure 10 directly compares the correct clustering rates of the above eight methods over categories. In Figure 10 each bar reflects the average fraction of the top-19 closest members of each task’s ordered similarity list that are within the same category as the task under test (there are 20 examples in each category; therefore, this measure examines the degree to which a given data is deemed most similar to all other data within the same class, before being similar to data from any other class). It is clear from Figure 10 that ordering based on atom information from both the image *and* text features yields best performance, and generally the MCMC and VB HLPP* implementations provide similar performance. Typical example sorting results are depicted in Figure 11. Considering Figure 10 more carefully, we note that in some cases using $\rho_k = 1$ for $k = 1, 2$ (Barnard et al., 2003; Blei and Jordan, 2003) yields significantly worse clustering performance than HLPP*, in particular for Building, Coast, Forest, Office and Bedroom. Additionally, in some cases using $\rho_k = 0$ for $k = 1, 2$ (treating the image and text features independently in the prior) and employing both of the posterior atom-weight vectors to measure similarity actually does better than directly coupling the mixture weights in the prior, in particular for Coast, Forest, Bedroom and Office, with this attributed to imperfections in the image features and in the annotations.

To demonstrate the prediction performance of the HLPP* model, Figure 12 reports the correct clustering rates for test data (now based on the average fraction of the top-10 closest members of the ranking list for each data); here we only show the top-10 (rather than 19) because these samples were *not* in the training set. From the results shown in Figure 12, we observe that MCMC and VB using the posterior atom-weight vectors over both image *and* annotation features achieved good test results for highway, building, coast and forest categories (larger than 0.85). Since the bedroom and office categories are somewhat confused in the training procedure, their test results are also not as good as those of other categories (about 0.75). Nevertheless, these are only clustering results; it is anticipated that one may design classifiers based on the posterior atom-weight vectors to improve the test performance (in future work). In addition, the results using the posterior atom-weight vectors over image features and annotation features *alone* are also good (see Figure 12). This demonstrates that our model may also be used for automatic annotation and text-based image retrieval.

The computation required for *learning* the models via MCMC is expensive: it required over 48 hours (for the burn-in and collection period given above), using non-optimized MatlabTM software on a Pentium IV PC with a 2.1 GHz CPU. By contrast, the VB solution is much faster, requiring less than 4 hours, with typically about 50 iterations needed for convergence. For both versions, the *testing* on new data in the sense discussed in Section 7.1

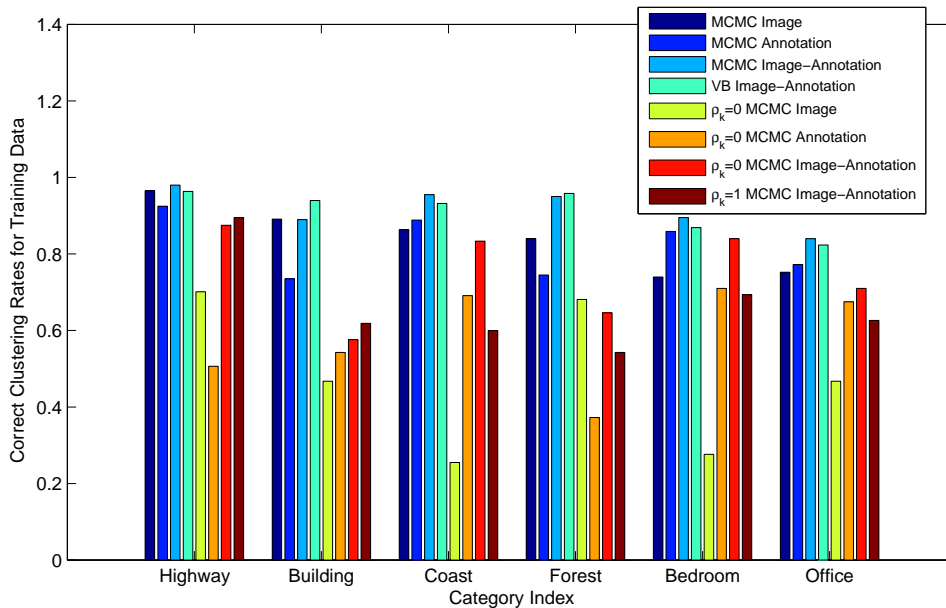


Figure 10: Each bar reflects the average fraction of the top-19 closest samples that are within the same category as the sample under test (there are 20 examples in each category). The ranking is based on the KL divergences on the posterior atom-weight vectors. Unless specified, the results are based on MCMC. The HLPP* results are computed via the posterior weights on image-feature atoms *alone*, annotation-features atoms *alone*, and both image-feature *and* annotation-feature atoms. The results in two limiting regimes of the HLPP* model with respect to $\{\rho_k\}_{k=1}^2$ are also shown, where $\rho_k = 0$ for $k = 1, 2$ corresponds to treating the image and text features independently in the prior; $\rho_k = 1$ for $k = 1, 2$ corresponds to the situation in (Barnard et al., 2003; Blei and Jordan, 2003)

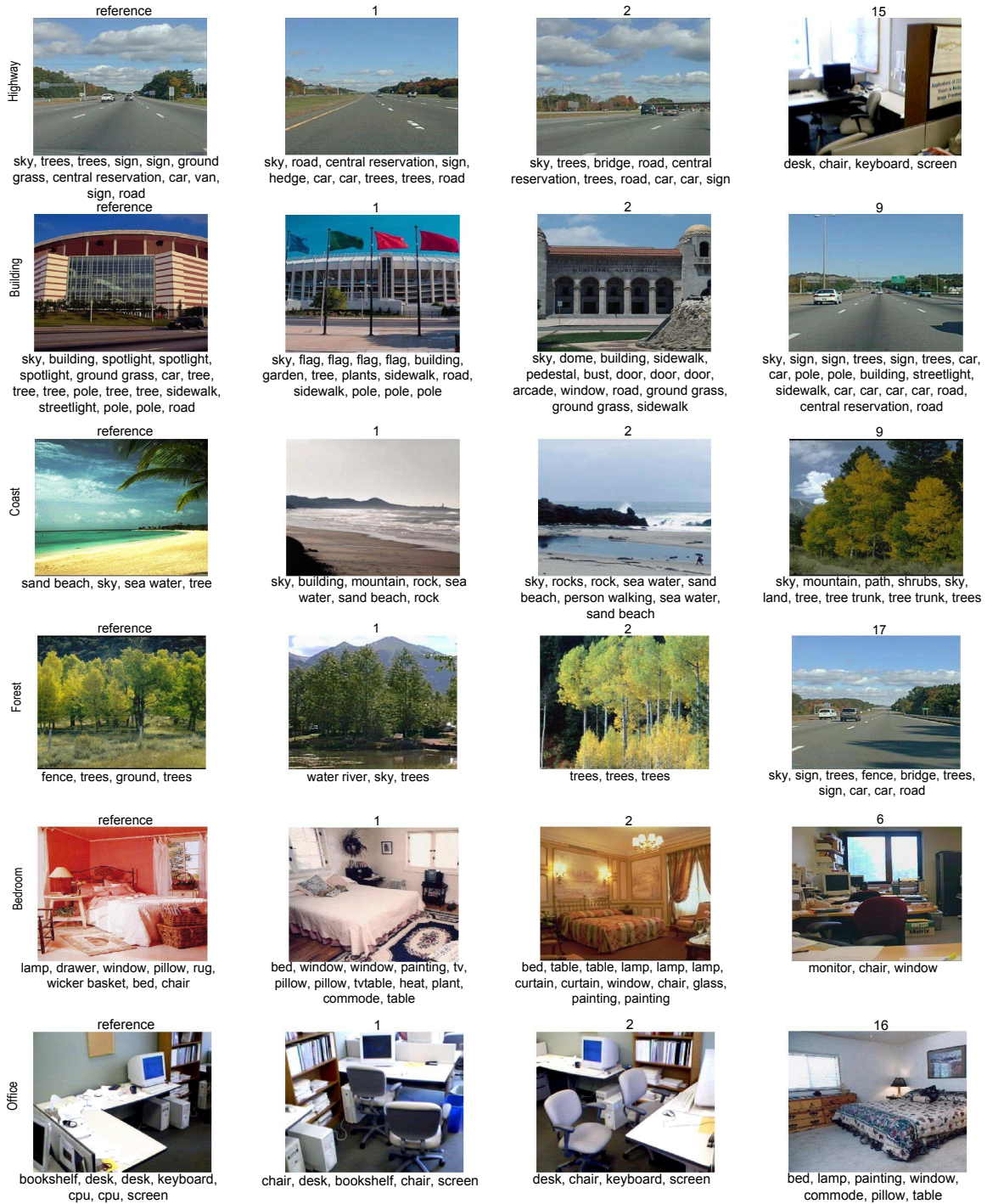


Figure 11: Typical example sorting results based on VB HLPP*; those for the MCMC inference are similar. The results are based on the posterior atom-weight vectors over both image *and* annotation features. Each row is for one category. The first column shows one template sample, and the next three columns show examples of ranked sorting results, with the number above the image specifying the closeness rank order as computed by the KL divergences. The last column denotes the first sample in the ranking that was in a wrong category.

is fast (a few minutes in non-optimized MatlabTM). This is important for online information retrieval application.

8. Conclusion

We have developed a new hierarchical Bayesian model for hybrid multi-task learning using the LPP prior (Dunson). The LPP prior can combine global and local clusterings through a locally-weighted mixture, and therefore it leads to dependent local clustering and borrowing of information among different modalities. In order to extend LPP to model multiple infinite mixture models, the DP-type local and global components in the original LPP model have been replaced with the HDP, yielding the proposed HLPP model. This model shares data across multiple modalities and multiple tasks. Inference is performed via the combination of collapsed and blocked Gibbs sampling and VB scheme.

The performance of HLPP model has been demonstrated using a simulated example, gene-expression data, and image-annotation analysis. Compared with independent HDP models for the different modalities and a completely dependent HDP model for combining the multiple modalities, the HLPP model has demonstrated improved clustering performance. In addition, the HLPP model can provide a smaller estimation uncertainty for most parameters, which is attributed to proper global and local sharing explicitly imposed within the HLPP prior. In addition, the developed VB solution allows relatively fast computation, which can be used for modeling and sorting large databases, such as the application of the image retrieval system.

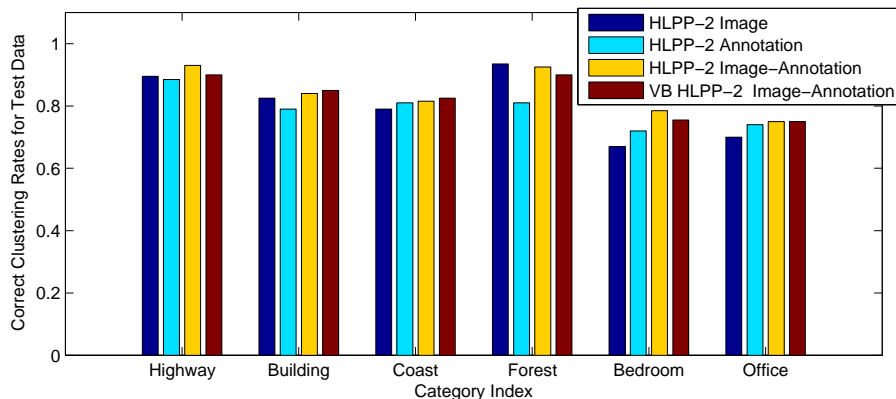


Figure 12: Using a separate set of data for testing the models, each bar reflects the average fraction of the top-10 closest samples that are within the same category as the sample under test (there are 20 examples in each category). The ranking is based on the KL divergences on the posterior atom-weight vectors. The results are computed via the HLPP* inferred posterior weights on image-feature atoms *alone*, annotation-feature atoms *alone*, and both the image-feature *and* annotation-feature atoms. The first three results are for MCMC inference, and the last one for VB.

Concerning future work, there may be other forms of prior information that one may wish to impose. For example, in the time-evolving gene-expression example, the different tasks corresponded to data collected at known time points. Typically one might expect the likelihood of task clustering to enhance as the tasks become more proximate. Related ideas have been considered recently for simpler DP-based models (Ren et al., 2008), and may also be considered in future work for LPP and HLPP. One may also be interested in addressing the realistic problem for which the data from a given task is incomplete, in that not all modalities have been employed for a given task. For the case in which such missing data is manifested completely randomly, the model presented here may be extended directly.

In addition, there are alternative means by which to implement the basic HLPP construction. For example, rather than employing an HDP prior for draws of $G^{(m)}$ and G_k , one may alternatively employ a nested Dirichlet process (Rodríguez et al., 2008). In such a construction the sharing between tasks is more explicit, since one need not always share the same set of atoms across tasks, as in the HDP.

Appendix A.

In this appendix we prove Property i) of the HLPP model from Section 3:

Proof. If data $\mathbf{x}_{l_3 k_2}^{(m_1)}$ and $\mathbf{x}_{l_4 k_2}^{(m_2)}$ are contained within the same cluster, we wish it more probable that $\mathbf{x}_{l_1 k_1}^{(m_1)}$ and $\mathbf{x}_{l_2 k_1}^{(m_2)}$ will be in the same cluster, for $k_1 \neq k_2$, $m_1 \neq m_2$, $l_1 \in \{1, 2, \dots, L_1^{(m_1)}\}$, $l_2 \in \{1, 2, \dots, L_1^{(m_2)}\}$, $l_3 \in \{1, 2, \dots, L_2^{(m_1)}\}$ and $l_4 \in \{1, 2, \dots, L_2^{(m_2)}\}$.

$$\begin{aligned}
\Pr(\boldsymbol{\theta}_{l_1 k_1}^{(m_1)} = \boldsymbol{\theta}_{l_2 k_1}^{(m_2)}) &= \int \sum_{i_1=0}^1 \sum_{i_2=0}^1 \Pr(\boldsymbol{\theta}_{l_1 k_1}^{(m_1)} = \boldsymbol{\theta}_{l_2 k_1}^{(m_2)} | z_{k_1}^{(m_1)} = i_1, z_{k_1}^{(m_2)} = i_2) \\
&\quad \cdot \Pr(z_{k_1}^{(m_1)} = i_1, z_{k_1}^{(m_2)} = i_2) \cdot (\rho_{k_1}) \, d\rho_{k_1} \\
&= \mathbb{E}\left[\sum_{j=1}^{\infty} v_{0j}^{(m_1)} v_{0j}^{(m_2)}\right] \cdot \mathbb{E}[\rho_{k_1}^2] + \mathbb{E}\left[\sum_{j=1}^{\infty} v_{k_1 j}^{(m_1)} v_{k_1 j}^{(m_2)}\right] \cdot \mathbb{E}[(1 - \rho_{k_1})^2] \\
&= \int \left\{ \sum_{j=1}^{\infty} \mathbb{E}^2[v_{0j}^{(m_1)} | \boldsymbol{\omega}'] \cdot \mathbb{E}[\rho_{k_1}^2] + \sum_{j=1}^{\infty} \mathbb{E}^2[v_{k_1 j}^{(m_1)} | \boldsymbol{\omega}'] \cdot \mathbb{E}[(1 - \rho_{k_1})^2] \right\} p(\boldsymbol{\omega}') \, d\boldsymbol{\omega}' \\
&= \mathbb{E}\left[\sum_{j=1}^{\infty} \omega_j'^2 \prod_{h=1}^{j-1} (1 - \omega_h')^2\right] \cdot (\mathbb{E}[\rho_{k_1}^2] + \mathbb{E}[(1 - \rho_{k_1})^2]) \\
&= \left(\frac{1}{1 + \alpha}\right) \left(\frac{1}{2 + \beta}\right) \left(\beta + \frac{2}{1 + \beta}\right) = \mathcal{P}_1(\alpha, \beta) \tag{40}
\end{aligned}$$

In the above derivation, since $\mathbf{v}_{k'}^{(m)}$ *i.i.d.* DP($\alpha, \boldsymbol{\omega}$), $\mathbb{E}[\mathbf{v}_{k'}^{(m)}|\boldsymbol{\omega}'] = \boldsymbol{\omega}$ with $\omega_j = \sum_{j=1}^{\infty} \omega'_j \prod_{h=1}^{j-1} (1 - \omega'_h)$, for $k' = 0, 1, \dots, K$ and $m = 1, 2, \dots, M$.

$$\begin{aligned}
\Pr(\boldsymbol{\theta}_{l_1 k_1}^{(m_1)} = \boldsymbol{\theta}_{l_2 k_1}^{(m_2)}, \boldsymbol{\theta}_{l_3 k_2}^{(m_1)} = \boldsymbol{\theta}_{l_4 k_2}^{(m_2)}) &= \mathbb{E}\left[\sum_{j_1=1}^{\infty} v_{0j_1}^{(m_1)} v_{0j_1}^{(m_2)} \sum_{j_2=1}^{\infty} v_{0j_2}^{(m_1)} v_{0j_2}^{(m_2)}\right] \cdot \mathbb{E}[\rho_{k_1}^2 \rho_{k_2}^2] \\
&+ \mathbb{E}\left[\sum_{j_1=1}^{\infty} v_{0j_1}^{(m_1)} v_{0j_1}^{(m_2)} \sum_{j_2=1}^{\infty} v_{k_2 j_2}^{(m_1)} v_{k_2 j_2}^{(m_2)}\right] \cdot \mathbb{E}[\rho_{k_1}^2 (1 - \rho_{k_2})^2] \\
&+ \mathbb{E}\left[\sum_{j_1=1}^{\infty} v_{k_1 j_1}^{(m_1)} v_{k_1 j_1}^{(m_2)} \sum_{j_2=1}^{\infty} v_{0j_2}^{(m_1)} v_{0j_2}^{(m_2)}\right] \cdot \mathbb{E}[(1 - \rho_{k_1})^2 \rho_{k_2}^2] \\
&+ \mathbb{E}\left[\sum_{j_1=1}^{\infty} v_{k_1 j_1}^{(m_1)} v_{k_1 j_1}^{(m_2)} \sum_{j_2=1}^{\infty} v_{k_2 j_2}^{(m_1)} v_{k_2 j_2}^{(m_2)}\right] \cdot \mathbb{E}[(1 - \rho_{k_1})^2 (1 - \rho_{k_2})^2] \\
&= \left\{ \sum_{j_1=1}^{\infty} \mathbb{E}\left[\left(\omega'_{j_1} + \text{Var}[v_{0j_1}^{(m_1)}|\boldsymbol{\omega}']\right)^2 \prod_{h_1=1}^{j_1-1} \left((1 - \omega'_{h_1})^2 + \text{Var}[v_{0h_1}^{(m_1)}|\boldsymbol{\omega}']\right)^2\right] \right. \\
&\quad \left. + \left(\sum_{j_1=1}^{\infty} \mathbb{E}\left[\omega'_{j_1}{}^2 \prod_{h_1=1}^{j_1-1} (1 - \omega'_{h_1})^2\right] \sum_{j_2: j_2 \neq j_1} \mathbb{E}\left[\omega'_{j_2}{}^2 \prod_{h_2=1}^{j_2-1} (1 - \omega_{h_2})^2\right] \right) \right\} \cdot \mathbb{E}^2[\rho_{k_1}^2] \\
&\quad + \mathbb{E}\left[\left(\sum_{j_1=1}^{\infty} \omega'_{j_1}{}^2 \prod_{h_1=1}^{j_1-1} (1 - \omega'_{h_1})^2\right)^2\right] \cdot \left\{ 2\mathbb{E}[\rho_{k_1}^2] \cdot \mathbb{E}[(1 - \rho_{k_1})^2] + \mathbb{E}^2[(1 - \rho_{k_1})^2] \right\} \\
&= \mathcal{P}_2(\alpha, \beta, \gamma) \\
&= \left\{ \mathbb{E}\left[\left(\sum_{j_1=1}^{\infty} \omega'_{j_1}{}^2 \prod_{h_1=1}^{j_1-1} (1 - \omega'_{h_1})^2\right)^2\right] \right. \\
&\quad \left. \cdot \left\{ \mathbb{E}^2[\rho_{k_1}^2] + 2\mathbb{E}[\rho_{k_1}^2] \cdot \mathbb{E}[(1 - \rho_{k_1})^2] + \mathbb{E}^2[(1 - \rho_{k_1})^2] \right\} \right\} \\
&\quad + \left\{ \left\{ \sum_{j_1=1}^{\infty} \mathbb{E}[\text{Var}^2[v_{0j_1}^{(m_1)}|\boldsymbol{\omega}'] \prod_{h_1=1}^{j_1-1} ((1 - \omega'_{h_1})^2 + \text{Var}[v_{0h_1}^{(m_1)}|\boldsymbol{\omega}'])^2 + 2\omega'_{j_1}{}^2 \text{Var}[v_{0j_1}^{(m_1)}|\boldsymbol{\omega}'] \right. \right. \\
&\quad \left. \left. \cdot \prod_{h_1=1}^{j_1-1} ((1 - \omega'_{h_1})^2 + \text{Var}[v_{0h_1}^{(m_1)}|\boldsymbol{\omega}'])^2 + \omega'_{j_1}{}^4 \prod_{h_1=1}^{j_1-1} \text{Var}^2[v_{0h_1}^{(m_1)}|\boldsymbol{\omega}']] \right\} \cdot \mathbb{E}^2[\rho_{k_1}^2] \right\} \\
&= \mathcal{P}_{2,1}(\alpha, \beta) + \mathcal{P}_{2,2}(\alpha, \beta, \gamma) \\
&= \frac{6 + \alpha}{(1 + \alpha)(2 + \alpha)(3 + \alpha)} \left[\frac{4}{(1 + \beta)^2(2 + \beta)^2} + \frac{2 \times 2}{(1 + \beta)(2 + \beta)} \left(\frac{\beta}{2 + \beta}\right) + \left(\frac{\beta}{2 + \beta}\right)^2 \right] \\
&\quad + \mathcal{P}_{2,3}(\alpha, \gamma) \frac{4}{(1 + \beta)^2(2 + \beta)^2} \\
&= \frac{6 + \alpha}{(1 + \alpha)(2 + \alpha)(3 + \alpha)(2 + \beta)^2} \left(\beta + \frac{2}{1 + \beta}\right)^2 + \frac{4\mathcal{P}_{2,3}(\alpha, \gamma)}{(1 + \beta)^2(2 + \beta)^2}
\end{aligned}$$

$$\begin{aligned}
\mathcal{P}_{2,1}(\alpha, \beta) &= \mathbb{E} \left[\left(\sum_{j_1=1}^{\infty} \omega'_{j_1}{}^2 \prod_{h_1=1}^{j_1-1} (1 - \omega'_{h_1})^2 \right)^2 \right] \\
&\quad \cdot \left\{ \mathbb{E}^2[\rho_{k_1}^2] + 2\mathbb{E}[\rho_{k_1}^2] \cdot \mathbb{E}[(1 - \rho_{k_1})^2] + \mathbb{E}^2[(1 - \rho_{k_1})^2] \right\} \\
\mathcal{P}_{2,2}(\alpha, \beta, \gamma) &= \mathcal{P}_{2,3}(\alpha, \gamma) \cdot \mathbb{E}^2[\rho_{k_1}^2] \\
\mathcal{P}_{2,3}(\alpha, \gamma) &= \sum_{j_1=1}^{\infty} \mathbb{E}[\text{Var}^2[v_{0j_1}^{(m_1)} | \boldsymbol{\omega}'] \prod_{h_1=1}^{j_1-1} ((1 - \omega'_{h_1})^2 + \text{Var}[v_{0h_1}^{(m_1)} | \boldsymbol{\omega}'])^2 + 2\omega'_{j_1}{}^2 \text{Var}[v_{0j_1}^{(m_1)} | \boldsymbol{\omega}'] \\
&\quad \cdot \prod_{h_1=1}^{j_1-1} ((1 - \omega'_{h_1})^2 + \text{Var}[v_{0h_1}^{(m_1)} | \boldsymbol{\omega}'])^2 + \omega'_{j_1}{}^4 \prod_{h_1=1}^{j_1-1} \text{Var}^2[v_{0h_1}^{(m_1)} | \boldsymbol{\omega}']] \\
\text{Var}[v_{0j_1}^{(m_1)} | \boldsymbol{\omega}'] &= \frac{\omega'_{j_1} - \omega'_{j_1}{}^2}{\gamma \prod_{h_1=1}^{j_1-1} (1 - \omega'_{h_1}) + 1}
\end{aligned}$$

Since $\text{Var}[v_{0j_1}^{(m_1)} | \boldsymbol{\omega}'] \geq 0$, $\mathcal{P}_{2,2}(\alpha, \beta, \gamma) \geq 0$ (while there is no analytical solution for $\mathcal{P}_{2,3}(\alpha, \gamma)$). We can easily prove $\frac{\mathcal{P}_{2,1}(\alpha, \beta)}{\mathcal{P}_1(\alpha, \beta)} \geq \mathcal{P}_1(\alpha, \beta)$, and then $\frac{\mathcal{P}_2(\alpha, \beta, \gamma)}{\mathcal{P}_1(\alpha, \beta)} \geq \mathcal{P}_1(\alpha, \beta)$, therefore, Property 1 in Section 3 is satisfied.

Appendix B.

In this appendix we prove the Property ii) of the HLPP model from Section 3:

Proof. Note that

$$\begin{aligned}
&\Pr \left(\left[\Pr(\boldsymbol{\theta}_{l_1 k_1}^{(m_1)} = \boldsymbol{\theta}_{l_2 k_1}^{(m_2)} | \boldsymbol{\theta}_{l_3 k_2}^{(m_1)} = \boldsymbol{\theta}_{l_4 k_2}^{(m_2)}) - \Pr(\boldsymbol{\theta}_{l_1 k_1}^{(m_1)} = \boldsymbol{\theta}_{l_2 k_1}^{(m_2)}) \right] \in S \right) \\
&= \int \mathbf{1}(\Delta(\alpha, \beta, \gamma) \in S) \mathcal{F}(\alpha, \beta, \gamma) d\alpha d\beta d\gamma
\end{aligned} \tag{41}$$

where $\Delta(\alpha, \beta, \gamma) = \mathcal{P}_2(\alpha, \beta, \gamma) / \mathcal{P}_1(\alpha, \beta) - \mathcal{P}_1(\alpha, \beta)$, and $\mathcal{F}(\alpha, \beta, \gamma)$ is the prior density for α , β and γ . Since we use independent gamma priors, $\mathcal{F}(\alpha, \beta, \gamma) > 0$ for all $(\alpha, \beta, \gamma) \in (0, \infty) \times (0, \infty) \times (0, \infty)$. For any point $\iota \in S$, there exists a corresponding region $\vartheta(\iota) \subset (0, \infty) \times (0, \infty) \times (0, \infty)$ of (α, β, γ) values that result in $\Delta(\alpha, \beta, \gamma) = \iota$, with $\vartheta(\iota) \neq \emptyset$ for all $\iota \in (0, 1)$. This condition follows directly if there exists an (α, β, γ) solution to the equation $\Delta(\alpha, \beta, \gamma) = \iota$, for every point ι in $0 < \iota < 1$. According to (40) and (??)

$$\Delta(\alpha, \beta, \gamma) = \frac{2\alpha}{(1+\alpha)(2+\alpha)(3+\alpha)} \frac{\beta^2 + \beta + 2}{(1+\beta)(2+\beta)} + \frac{4(1+\alpha)\mathcal{P}_{2,3}(\alpha, \gamma)}{(1+\beta)(2+\beta)(\beta^2 + \beta + 2)} \tag{42}$$

If $\alpha \rightarrow 0$ and $\gamma \rightarrow \infty$, then $\Delta(0, \beta, \infty) \rightarrow 0$. When $\alpha \rightarrow \infty$ and $\beta \rightarrow 0$, we can simplify the HLPP model and easily obtain

$$\Pr(\boldsymbol{\theta}_{l_1 k_1}^{(m_1)} = \boldsymbol{\theta}_{l_2 k_1}^{(m_2)}) \rightarrow 0, \quad \Pr(\boldsymbol{\theta}_{l_1 k_1}^{(m_1)} = \boldsymbol{\theta}_{l_2 k_1}^{(m_2)} | \boldsymbol{\theta}_{l_3 k_2}^{(m_1)} = \boldsymbol{\theta}_{l_4 k_2}^{(m_2)}) \rightarrow \frac{6+\gamma}{(2+\gamma)(3+\gamma)} \tag{43}$$

Thus $\Delta(\infty, 0, \gamma) \rightarrow (6+\gamma) / [(2+\gamma)(3+\gamma)]$. If $\gamma \rightarrow 0$, then $\Delta(\infty, 0, 0) \rightarrow 1$. Since $\Delta(\alpha, \beta, \gamma)$ should be a continuous function, $\Pr(\Delta(\alpha, \beta, \gamma) \in S) > 0$ for all Borel subset $S \subset [0, 1]$.

References

- E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.
- Q. An, C. Wang, I. Shterev, E. Wang, D. B. Dunson, and L. Carin. Hierarchical kernel stick-breaking process for multi-task image analysis. In *International Conference of Machine Learning (ICML)*, 2008.
- K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. M. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.
- M. J. Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.
- D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
- D. M. Blei and M. I. Jordan. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 127–134, 2003.
- E. Brochu, N. de Freitas, and K. Bao. The sound of an album cover: Probabilistic multimedia and ir. In *Ninth International Workshop on Artificial Intelligence and Statistics*, 2003. URL <http://www.cs.ubc.ca/nando/papers/aismedia.pdf>.
- C. Carvalho, J. Chang, J. Lucas, J. R. Nevins, Q. Wang, and M. West. High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association*.
- J. T. Chang and J. R. Nevins. Sgather: a systems approach to interpreting genomic signatures. *Bioinformatics*, 22(23):2926–2933, 2006.
- P. Common. Independent component analysis - a new concept? *Signal Processing*, 36:287–314, 1994.
- D. B. Dunson. Nonparametric bayes local partition models for random effects. *Biometrika*.
- D. B. Dunson, Y. Xue, and L. Carin. The matrix stick-breaking process: Flexible bayes meta analysis. *Journal of the American Statistical Association*, 103:317–327, 2008.
- J. Fink, F. Gu, L. Ling, T. Tolfvenstan, F. Olfat, K. C. Chin, P. Aw, J. George, V. A. Kuznetsov, M. Schreiber, S. G. Vasudevan, and M. L. Hibberd. Host gene expression profiling of dengue virus infection in cell lines and patients. *PLoS Neglected Tropical Diseases*, 1(2), 2007.
- E. Fokoue. Stochastic determination of the intrinsic structure in bayesian factor analysis. Technical report, 2004.
- J. Geweke. Evaluating the accuracy of sampling-based approaches to calculation of posterior moments. *Bayesian Statistics*, 4:169–193, 1992.

- W. R. Gilks and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. London: Chapman Hall, 1996.
- T. L. Griffiths and Z. Ghahramani. Infinite latent feature models and the indian buffet process. Technical report, 2005. URL <http://cocosci.berkeley.edu/tom/papers/ibptr.pdf>.
- P. Hoyer and A. Hyvärinen. Independent component analysis applied to feature extraction from colour and stereo images. *Network: Computation in Neural Systems*, 11(3):191–210, 2000.
- A. Hyvärinen and P. Hoyer. Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspace. *Neural Computation*, 12(7):1705–1720, 2000.
- H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96:161173, 2001.
- J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 119–126, 2003.
- C. Kemp, T. L. Griffiths, and J. B. Tenenbaum. Discovering latent classes in relational data. Technical report, 2004.
- C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *Proceedings of the 21st National Conference on Artificial Intelligence*, 2006.
- D. Knowles and Z. Ghahramani. Infinite sparse factor analysis and infinite independent components analysis. In *7th International Conference on Independent Component Analysis and Signal Separation (ICA)*, 2007.
- K. Livescu, K. Sridharan, S. Kakade, and K. Chauduri. Multi-view clustering via canonical correlation analysis. In *Neural Information Processing Systems Conference*, 2008.
- S. N. MacEachern. Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics*, 7:223–238, 1998.
- P. Muliere and L. Tardella. Approximating distributions of random functionals of ferguson-dirichlet priors. *Can. J. Statist.*, 26:283–97, 1998.
- S. Petrone, M. Guindani, and A. E. Gelfand. Hybrid dirichlet processes for functional data. In *An Isaac Newton Institute Workshop: Construction and Proferties of Bayesian Nonparametric Regression Models*, 2008. URL <http://www.newton.ac.uk/programmes/BNR/seminars/080811301.html>.
- I. Pournara and L. Wernish. Factor analysis for gene regulatory networks and transcription factor activity profiles. *BMC Bioinformatics*, 8, 2007.

- A. E. Raftery and S. Lewis. How many iterations in the gibbs sampler? *Bayesian Statistics*, 4:763–773, 1992.
- L. Ren, D. B. Dunson, and L. Carin. The dynamic hierarchical dirichlet process. In *International Conference of Machine Learning (ICML)*, 2008.
- A. Rodríguez, D. B. Dunson, and A. E. Gelfang. The nested dirichlet process. *Journal of the American Statistical Association*, 103:1131–1144, 2008.
- B. Taskar, E. Segal, and D. Koller. Probabilistic clustering in relational data. In *IJCAI-01*, pages 870–876, 2001.
- Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.
- Y. W. Teh, K. Kurihara, and M. Welling. Collapsed variational inference for hdp. In *Neural Information Processing Systems Conference*, 2007. URL <http://www.ics.uci.edu/welling/publications/papers/CVHDP.pdf>.
- S. G. Walker. Sampling the dirichlet mixture model with slices. *Comm. Statist. Sim. Comput.*, 36:45–54, 2007.
- Y. J. Wang and G. Y. Wong. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82:819, 1987.
- F. Wood, T. L. Griffiths, and Z. Ghahramani. A non-parametric bayesian method for inferring hidden causes. In *Uncertainty in Artificial Intelligence (UAI)*, pages 536–543, 2006.
- Z. Xu, V. Tresp, K. Yu, and H. Kriegel. Learning infinite hidden relational models. In *Uncertainty in Artificial Intelligence (UAI)*, 2006.
- Y. Xue, X. Liao, L. Carin, and B. Krishnapuram. Multi-task learning for classification with dirichlet process priors. *Journal of Machine Learning Research*, 8:35–63, 2007.