Diversity-Promoting Bayesian Learning of Latent Variable Models

Pengtao Xie[†] PENGTAOX@CS.CMU.EDU
Jun Zhu[†‡] DCSZJ@TSINGHUA.EDU.CN
Eric P. Xing[†] EPXING@CS.CMU.EDU
[†]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213 USA
[‡]Dept. of Comp. Sci. & Tech., State Key Lab of Intell. Tech. & Sys., TNList, CBICR Center, Tsinghua University, China

Abstract

In learning latent variable models (LVMs), it is important to effectively capture infrequent patterns and shrink model size without sacrificing modeling power. Various studies have been done to "diversify" a LVM, which aim to learn a diverse set of latent components in LVMs. Most existing studies fall into a frequentist-style regularization framework, where the components are learned via point estimation. In this paper, we investigate how to "diversify" LVMs in the paradigm of Bayesian learning, which has advantages complementary to point estimation, such as alleviating overfitting via model averaging and quantifying uncertainty. We propose two approaches that have complementary advantages. One is to define diversity-promoting mutual angular priors which assign larger density to components with larger mutual angles based on Bayesian network and von Mises-Fisher distribution and use these priors to affect the posterior via Bayes rule. We develop two efficient approximate posterior inference algorithms based on variational inference and Markov chain Monte Carlo sampling. The other approach is to impose diversity-promoting regularization directly over the post-data distribution of components. These two methods are applied to the Bayesian mixture of experts model to encourage the "experts" to be diverse and experimental results demonstrate the effectiveness and efficiency of our methods.

1. Introduction

Latent variable models (LVMs) (Bishop, 1998; Knott & Bartholomew, 1999; Blei, 2014) are a major workhorse in machine learning (ML) to extract latent *patterns* underlying data, such as *themes* behind documents and *motifs* hidden in genome sequences. To properly capture these patterns, LVMs are equipped with a set of *components*, each of which is aimed to capture one pattern and is usually parametrized by a vector. For instance, in topic models (Blei et al., 2003), each component (referred to as *topic*) is in charge of capturing one *theme* underlying documents and is represented by a multinomial vector.

While existing LVMs have demonstrated great success, they are less capable in addressing two new problems emerged due to the growing volume and complexity of data. First, it is often the case that the frequency of patterns is distributed in a power-law fashion (Wang et al., 2014; Xie et al., 2015) where a handful of patterns occur very frequently whereas most patterns are of low frequency (Figure 1 shows an example). Existing LVMs lack capability to capture infrequent patterns, which is possibly due to the design of LVMs' objective function used for training. For example, a maximum likelihood estimator would reward itself by modeling the frequent patterns well as they are the major contributors of the likelihood function. On the other hand, infrequent patterns contribute much less to the likelihood, thereby it is not very rewarding to model them well and LVMs tend to ignore them. Infrequent patterns often carry valuable information, thus should not be ignored. For instance, in a topic modeling based recommendation system, an infrequent topic (pattern) like *losing weight* is more likely to improve the click-through rate than a frequent topic like *politics*. Second, the number of components K strikes a tradeoff between model size (complexity) and modeling power. For a small K, the model is not expressive enough to sufficiently capture the complex patterns behind data; for a large K, the model would be of large size and complexity, incurring high computational overhead. How to reduce model size while preserving modeling power is a challenging issue.

To cope with the two problems, several studies (Zou & Adams, 2012; Xie et al., 2015; Xie, 2015) propose a "diver-

Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

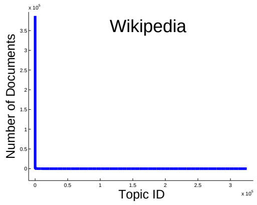

Figure 1. Power-law distribution of topic frequency (measured by number of documents in each topic) in the Wikipedia dataset.

sification" approach, which encourages the components of a LVM to be mutually "dissimilar". First, regarding capturing infrequent patters, as posited in (Xie et al., 2015) "diversified" components are expected to be less aggregated over frequent patterns and part of them would be spared to cover the infrequent patterns. Second, concerning shrinking model size without compromising modeling power, Xie (2015) argued that "diversified" components bear less redundancy and are mutually complementary, making it possible to capture information sufficiently well with a small set of components, i.e., obtaining LVMs possessing high representational power and low computational complexity.

The existing studies (Zou & Adams, 2012; Xie et al., 2015; Xie, 2015) of "diversifying" LVMs mostly focus on *point estimation* (Wasserman, 2013) of the model components, under a frequentist-style regularized optimization framework. In this paper, we study how to promote diversity under an alternative learning paradigm: Bayesian inference (Jaakkola & Jordan, 1997; Bishop & Tipping, 2003; Neal, 2012), where the components are considered as random variables of which a *posterior distribution* shall be computed from data under certain priors. Compared with point estimation, Bayesian learning offers complementary benefits. First, it offers a "model-averaging" (Jaakkola & Jordan, 1997; Bishop & Tipping, 2003) effect for LVMs when they are used for decision-making and prediction because the parameters shall be integrated under a posterior distribution, thus potentially alleviate overfitting on training data. Second, it provides a natural way to quantify uncertainties of model parameters, and downstream decisions and predictions made thereupon (Jaakkola & Jordan, 1997; Bishop & Tipping, 2003; Neal, 2012). Affandi et al. (2013) investigated the "diversification" of Bayesian LVMs using the determinantal point process (DPP) (Kulesza & Taskar, 2012) prior. While Markov chain Monte Carlo (MCMC) (Affandi et al., 2013) methods have been developed for approximate posterior inference under the DPP prior, DPP is not amenable for another mainstream paradigm of approximate inference techniques – variational inference (Wainwright & Jordan, 2008) – which is usually more efficient (Hoffman et al., 2013) than MCMC. In this paper, we propose alternative diversity-promoting priors that overcome this limitation.

We propose two approaches that have complementary advantages to perform diversity-promoting Bayesian learning of LVMs. Following (Xie et al., 2015), we adopt a notion of diversity that component vectors are more diverse provided the pairwise angles between them are larger. First, we define mutual angular Bayesian network (MABN) priors over the components, which assign higher probability density to components that have larger mutual angles and use these priors to affect the posterior via Bayes rule. Specifically, we build a Bayesian network (Koller & Friedman, 2009) where nodes represent the directional vectors of the components and local probabilities are parameterized by von Mises-Fisher (Mardia & Jupp, 2009) distributions which entail an inductive bias towards vectors with larger mutual angles. The MABN priors are amenable for approximate posterior inference of model components. In particular, they facilitate variational inference, which is usually more efficient than MCMC sampling. Second, in light of that it is not flexible (or even possible) to define priors to capture certain diversity-promoting effects such as small variance of mutual angles, we adopt a posterior regularization approach (Zhu et al., 2014), in which a diversity-promoting regularizer is directly imposed over the post-data distributions to encourage diversity and the regularizer can be flexibly defined to accommodate various desired diversity-promoting goals. We instantiate the two approaches to Bayesian mixture of experts model (BMEM) (Waterhouse et al., 1996) and experiments demonstrate the effectiveness and efficiency of our approaches.

Related Works Recent works (Zou & Adams, 2012; Xie et al., 2015; Xie, 2015) have studied the diversification of components in LVMs under a point estimation framework. In a multi-class classification problem, Malkin & Bilmes (2008) proposed to use the determinant of a covariance matrix to encourage classifiers to be different from each other. Zou & Adams (2012) leveraged the determinantal point process (DPP) (Kulesza & Taskar, 2012) to promote diversity in latent variable models. Xie et al. (2015) proposed a mutual angular regularizer that encourages model components to be mutually different where the dissimilarity is measured by angles.

Diversity-promoting Bayesian learning of LVMs has been investigated in (Affandi et al., 2013), which utilizes the DPP prior to induce bias towards diverse components. Affandi et al. (2013) developed a Gibbs sampling (Gilks, 2005) algorithm. But the determinant in DPP makes variational inference based algorithms very difficult to derive.

Contributions The major contributions of this work are:

- We propose mutual angular Bayesian network (MABN) priors which are biased towards components having large mutual angles, to promote diversity in Bayesian LVMs.
- We develop an efficient variational inference method

for posterior inference of model components under the MABN priors.

- To flexibly accommodate various diversity-promoting effects, we study a posterior regularization approach which directly imposes diversity-promoting regularization over the post-data distributions.
- Using Bayesian mixture of experts model as a study case, we empirically demonstrate the effectiveness and efficiency of our methods.

The rest of the paper is organized as follows. In Section 2, we introduce how to promote diversity in Bayesian LVMs. Section 3 gives experimental results and Section 4 concludes the paper.

2. Methods

In this section, we study the diversity-promoting Bayesian learning of latent variable models and investigate two approaches: (1) *prior control*, which defines diversity-promoting priors and uses them to affect the posterior via Bayes rule; (2) *posterior regularization*, which directly performs diversity-promoting regularization over post-data distributions. These two approaches have complementary advantages which will be discussed in detail below.

2.1. Diversity-Promoting Mutual Angular Prior

The first approach we take is to define priors which have inductive bias towards components that are more "diverse" and use them to affect the posterior via Bayes rule. We refer to this approach as *prior control*. While diversity can be defined in various ways, following (Xie et al., 2015) we adopt the notion that a set of component vectors are deemed to be more diverse if the pairwise angles between them are larger. We desire the priors to have two traits. First, to favor diversity, they assign a higher density to components having larger mutual angles. Second, the priors should facilitate posterior inference. In Bayesian learning, the easiness of posterior inference relies heavily on the prior (Blei & Lafferty, 2006; Wang & Blei, 2013).

One possible solution is to turn the mutual angular regularizer $\Omega(\mathbf{A})$ (Xie et al., 2015) that encourages a set of component vectors $\mathbf{A} = \{\mathbf{a}_i\}_{i=1}^K$ to have large mutual angles into a distribution $p(\mathbf{A}) = \frac{1}{Z}\exp(\Omega(\mathbf{A}))$ based on Gibbs measure (Kindermann et al., 1980), where Z is the partition function guaranteeing $p(\mathbf{A})$ integrates to one. The concern is that it is not sure whether $Z = \int_{\mathbf{A}} \exp(\Omega(\mathbf{A}))\mathrm{d}\mathbf{A}$ is finite, i.e., whether $p(\mathbf{A})$ is proper. When an improper prior is utilized in Bayesian learning, the posterior is also highly likely to be improper, except in a few special cases (Wasserman, 2013). Performing inference on improper posteriors is problematic.

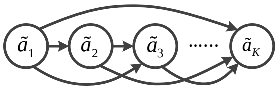

Figure 2. A Bayesian Network Representation of the Mutual Angular Prior

Here we define mutual angular Bayesian network (MABN) priors possessing the aforementioned two traits, based on Bayesian network (Koller & Friedman, 2009) and von Mises-Fisher (Mardia & Jupp, 2009) distribution. For technical convenience, we decompose each real-valued component vector $\mathbf{a}$ into $\mathbf{a} = g\tilde{\mathbf{a}}$, where $g = \|\mathbf{a}\|_2$ is the magnitude and $\tilde{\mathbf{a}}$ is the direction ($\|\tilde{\mathbf{a}}\|_2 = 1$). Let $\widetilde{\mathbf{A}} = \{\tilde{\mathbf{a}}_i\}_{i=1}^K$ denote the directional vectors. Note that the angle between two vectors is invariant to their magnitudes, thereby, the mutual angles of component vectors in $\mathbf{A}$ are the same as angles of directional vectors in $\widetilde{\mathbf{A}}$. We first construct a prior which prefers vectors in $\widetilde{\mathbf{A}}$ to possess large angles. The basic idea is to use a Bayesian network (BN) to characterize the dependency among directional vectors and design local probabilities to entail inductive bias towards large mutual angles. In the Bayesian network (BN) shown in Figure 2, each node i represents a directional vector $\tilde{\mathbf{a}}_i$ and its parents $\mathrm{pa}(\tilde{\mathbf{a}}_i)$ are nodes $1, \cdots, i-1$. We define local probability at node i to encourage $\tilde{\mathbf{a}}_i$ to have large mutual angles with $\tilde{\mathbf{a}}_1, \cdots, \tilde{\mathbf{a}}_{i-1}$. Since these directional vectors lie on a sphere, we use the von Mises-Fisher (vMF) distribution to model them. The probability density function of the vMF distribution is $f(\mathbf{x}) = C_p(\kappa)\exp(\kappa\boldsymbol{\mu}^\top\mathbf{x})$, where the random variable $\mathbf{x} \in \mathbb{R}^p$ lies on a $p-1$ dimensional sphere ($\|\mathbf{x}\|_2 = 1$), $\boldsymbol{\mu}$ is the mean direction with $\|\boldsymbol{\mu}\|_2 = 1$, $\kappa > 0$ is the concentration parameter and $C_p(\kappa)$ is the normalization constant. The local probability $p(\tilde{\mathbf{a}}_i|\mathrm{pa}(\tilde{\mathbf{a}}_i))$ at node i is defined as a von Mises-Fisher (vMF) distribution whose density is

$$p(\tilde{\mathbf{a}}_i|\mathrm{pa}(\tilde{\mathbf{a}}_i)) = C_p(\kappa)\exp\left(\kappa(-\frac{\sum_{j=1}^{i-1}\tilde{\mathbf{a}}_j}{\|\sum_{j=1}^{i-1}\tilde{\mathbf{a}}_j\|_2})^\top\tilde{\mathbf{a}}_i\right) \quad (1)$$

with mean direction $-\sum_{j=1}^{i-1}\tilde{\mathbf{a}}_j/\|\sum_{j=1}^{i-1}\tilde{\mathbf{a}}_j\|_2$. Now we explain why this local probability favors large mutual angles. Since $\tilde{\mathbf{a}}_i$ and $\tilde{\mathbf{a}}_j$ are unit-length vectors, $\tilde{\mathbf{a}}_j^\top\tilde{\mathbf{a}}_i$ is the cosine of the angle between $\tilde{\mathbf{a}}_i$ and $\tilde{\mathbf{a}}_j$. If $\tilde{\mathbf{a}}_i$ has larger angles with $\{\tilde{\mathbf{a}}_j\}_{j=1}^{i-1}$, then the average negative cosine similarity $(-\sum_{j=1}^{i-1}\tilde{\mathbf{a}}_j)^\top\tilde{\mathbf{a}}_i$ would be larger, accordingly $p(\tilde{\mathbf{a}}_i|\mathrm{pa}(\tilde{\mathbf{a}}_i))$ would be larger. This statement is true for all $i > 1$. As a result, $p(\widetilde{\mathbf{A}}) = p(\tilde{\mathbf{a}}_1)\prod_{i=2}^K p(\tilde{\mathbf{a}}_i|\mathrm{pa}(\tilde{\mathbf{a}}_i))$ would be larger if the directional vectors have larger mutual angles. For the magnitudes $\{g_i\}_{i=1}^K$ of the components, which have nothing to do with the mutual angles, we sample g_i for each component independently from a gamma distribution with shape parameter α_1 and rate parameter α_2. The generative process of $\mathbf{A}$ is summarized as follows:

- Draw $\tilde{\mathbf{a}}_1 \sim \text{vMF}(\boldsymbol{\mu}_0, \kappa)$
- For $i = 2, \cdots, K$, draw $\tilde{\mathbf{a}}_i \sim \text{vMF}(-\frac{\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j}{\|\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j\|_2}, \kappa)$
- For $i = 1, \cdots, K$, draw $g_i \sim \text{Gamma}(\alpha_1, \alpha_2)$
- For $i = 1, \cdots, K$, $\mathbf{a}_i = \tilde{\mathbf{a}}_i g_i$

The probability distribution over $\mathbf{A}$ can be written as

$$\begin{aligned} &p(\mathbf{A}) = C_p(\kappa) \exp(\kappa \mu_0^\top \tilde{\mathbf{a}}_1) \prod_{i=2}^K C_p(\kappa) \\ &\exp(\kappa(-\frac{\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j}{\|\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j\|_2})^\top \tilde{\mathbf{a}}_i) \prod_{i=1}^K \frac{\alpha_2^{\alpha_1} g_i^{\alpha_1 - 1} e^{-g_i \alpha_2}}{\Gamma(\alpha_1)} \end{aligned} \tag{2}$$

According to the factorization theorem (Koller & Friedman, 2009) of Bayesian network, it is easy to verify $\int_{\mathbf{A}} p(\mathbf{A}) \mathrm{d}\mathbf{A} = 1$, thus $p(\mathbf{A})$ is a proper prior.

When inferring the posterior of model components using a variational inference method, we need to compute the expectation of $1/\|\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j\|_2$ appearing in the local probability $p(\tilde{\mathbf{a}}_i|\text{pa}(\tilde{\mathbf{a}}_i))$, which is extremely difficult. To address this issue, we define an alternative local probability that achieves similar modeling effect as $p(\tilde{\mathbf{a}}_i|\text{pa}(\tilde{\mathbf{a}}_i))$, but greatly facilitates variational inference. We re-parametrize the local probability $\hat{p}(\tilde{\mathbf{a}}_i|\text{pa}(\tilde{\mathbf{a}}_i))$ defined in Eq.(1) using Gibbs measure:

$$\begin{aligned} &\hat{p}(\tilde{\mathbf{a}}_i|\text{pa}(\tilde{\mathbf{a}}_i)) \propto \exp(\kappa(-\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j)^\top \tilde{\mathbf{a}}_i) \\ &\propto \exp(\kappa\|\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j\|_2 (-\frac{\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j}{\|\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j\|_2})^\top \tilde{\mathbf{a}}_i) \\ &= C_p(\kappa\|\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j\|_2) \exp(\kappa\|\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j\|_2 (-\frac{\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j}{\|\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j\|_2})^\top \tilde{\mathbf{a}}_i) \\ &= C_p(\kappa\|\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j\|_2) \exp(\kappa(-\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j)^\top \tilde{\mathbf{a}}_i) \end{aligned} \tag{3}$$

which is another vMF distribution with mean direction $-\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j / \|\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j\|_2$ and concentration parameter $\kappa\|\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j\|_2$. This reparameterized local probability is proportional to $(-\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j)^\top \tilde{\mathbf{a}}_i$, which measures the negative cosine similarity between $\tilde{\mathbf{a}}_i$ and its parent vectors. Thereby, $\hat{p}(\tilde{\mathbf{a}}_i|\text{pa}(\tilde{\mathbf{a}}_i))$ still encourages large mutual angles between vectors as $p(\tilde{\mathbf{a}}_i|\text{pa}(\tilde{\mathbf{a}}_i))$ does. The difference between $\hat{p}(\tilde{\mathbf{a}}_i|\text{pa}(\tilde{\mathbf{a}}_i))$ and $p(\tilde{\mathbf{a}}_i|\text{pa}(\tilde{\mathbf{a}}_i))$ is that in $\hat{p}(\tilde{\mathbf{a}}_i|\text{pa}(\tilde{\mathbf{a}}_i))$ the term $\|\sum_{j=1}^{i-1} \tilde{a}_j\|_2$ is moved from the denominator to the normalizer, thus we can avoid computing the expectation of $1/\|\sum_{j=1}^{i-1} \tilde{a}_j\|_2$. Though it incurs a new problem that we need to compute the expectation of $\log C_p(\kappa\|\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j\|_2)$, which is also hard due to the complex form of the $C_p(\cdot)$ function, we managed to resolve this problem as detailed in Section 2.1.1. We refer to the MABN prior defined in Eq.(2) as type I MABN and that with local probability defined in Eq.(3) as type II MABN.

2.1.1. Approximate Inference Algorithms

We develop algorithms to infer the posteriors of components under the MABN priors. Since exact posteriors are intractable, we resort to approximate inference techniques. Two main paradigms of approximate inference methods are: (1) variational inference (VI) (Wainwright & Jordan, 2008); (2) Markov chain Monte Carlo (MCMC) sampling (Gilks, 2005). These two approaches possess benefits that are mutually complementary. MCMC can achieve a better approximation of the posterior than VI since it generates samples from the exact posterior while VI seeks an approximation. However, VI can be computationally more efficient (Hoffman et al., 2013).

Variational Inference The basic idea of VI (Wainwright & Jordan, 2008) is to use a "simpler" variational distribution $q(\mathbf{A})$ to approximate the true posterior by minimizing the Kullback-Leibler divergence between these two distributions, which is equivalent to maximizing the following variational lower bound w.r.t $q(\mathbf{A})$:

$$\mathbb{E}_{q(\mathbf{A})}[\log p(\mathcal{D}|\mathbf{A})] + \mathbb{E}_{q(\mathbf{A})}[\log p(\mathbf{A})] - \mathbb{E}_{q(\mathbf{A})}[\log q(\mathbf{A})] \tag{4}$$

where $p(\mathbf{A})$ is the MABN prior and $p(\mathcal{D}|\mathbf{A})$ is data likelihood. Here we choose $q(\mathbf{A})$ to be a mean field variational distribution $q(\mathbf{A}) = \prod_{k=1}^K q(\tilde{\mathbf{a}}_k) q(g_k)$, where $q(\tilde{\mathbf{a}}_k) = \text{vMF}(\tilde{\mathbf{a}}_k|\hat{\mathbf{a}}_k, \hat{\kappa})$ and $q(g_k) = \text{Gamma}(g_k|r_k, s_k)$. Given the variational distribution, we first compute the analytical expression of the variational lower bound, in which we particularly discuss how to compute $\mathbb{E}_{q(\mathbf{A})}[\log p(\mathbf{A})]$. If choosing $p(\mathbf{A})$ to be type I MABN prior (Eq.(2)), we need to compute $\mathbb{E}[(-\frac{\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j}{\|\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j\|_2})^\top \tilde{\mathbf{a}}_i]$ which is very difficult to deal with due to the presence of $1/\|\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j\|_2$. Instead we choose type II MABN prior for the convenience of deriving the variational lower bound. Under type II MABN, we need to compute $\mathbb{E}_{q(\mathbf{A})}[\log Z_i]$ for all i, where $Z_i = 1/C_p(\kappa\|\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j\|_2)$ is the partition function of $p(\tilde{\mathbf{a}}_i|\text{pa}(\tilde{\mathbf{a}}_i))$. The analytical form of this expectation is difficult to derive as well due to the complexity of the $C_p(x)$ function: $C_p(x) = \frac{x^{p/2-1}}{(2\pi)^{p/2} I_{p/2-1}(x)}$ where $I_{p/2-1}(x)$ is the modified Bessel function of the first kind at order $p/2 - 1$. To address this issue, we derive an upper bound of $\log Z_i$ and compute the expectation of the upper bound, which is relatively easy to do. Consequently, we obtain a further lower bound of the variational lower bound and learn the variational and model parameters w.r.t the new lower bound. Now we proceed to derive the upper bound of $\log Z_i$, which equals to $\log \int \exp(\kappa(-\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j) \cdot \tilde{\mathbf{a}}_i) \mathrm{d}\tilde{\mathbf{a}}_i$. Applying the inequality $\log \int \exp(x) \mathrm{d}x \leq \gamma + \int \log(1 + \exp(x - \gamma)) \mathrm{d}x$ (Bouchard, 2007), where γ is a variational parameter, we have

$$\log Z_i \leq \gamma + \int \log(1 + \exp(\kappa(-\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j) \cdot \tilde{\mathbf{a}}_i - \gamma) \mathrm{d}\tilde{\mathbf{a}}_i \tag{5}$$

Then applying the inequality $\log(1+e^{-x}) \leq \log(1+e^{-\xi}) - \frac{x-\xi}{2} - \frac{1/2 - g(\xi)}{2\xi}(x^2 - \xi^2)$ (Bouchard, 2007), where ξ is another variational parameter and $g(\xi) = 1/(1+\exp(-\xi))$, we have

$$\begin{aligned}\log Z_i \leq \gamma + \textstyle\int [\log(1+e^{-\xi}) - \frac{\kappa(\sum\limits_{j=1}^{i-1}\tilde{\mathbf{a}}_j)\cdot\tilde{\mathbf{a}}_i+\gamma-\xi}{2} \\ -\tfrac{1/2-g(\xi)}{2\xi}((\kappa(\sum_{j=1}^{i-1}\tilde{\mathbf{a}}_j)\cdot\tilde{\mathbf{a}}_i+\gamma)^2-\xi^2)]\mathrm{d}\tilde{\mathbf{a}}_i\end{aligned} \tag{6}$$

Finally applying the following integrals on a high-dimensional sphere: (1) $\int_{\|\mathbf{y}\|_2=1} 1\mathrm{d}\mathbf{y} = \frac{2\pi^{(p+1)/2}}{\Gamma(\frac{p+1}{2})}$, (2) $\int_{\|\mathbf{y}\|_2=1} \mathbf{x}^\top\mathbf{y}\mathrm{d}\mathbf{y} = 0$, (3) $\int_{\|\mathbf{y}\|_2=1}(\mathbf{x}^\top\mathbf{y})^2\mathrm{d}\mathbf{y} \leq \|\mathbf{x}\|_2^2\frac{2\pi^{(p+1)/2}}{\Gamma(\frac{p+1}{2})}$, we get

$$\begin{aligned}\log Z_i \leq -\tfrac{1/2-g(\xi)}{2\xi}\kappa^2\|\sum_{j=1}^{i-1}\tilde{\mathbf{a}}_j\|_2^2\tfrac{2\pi^{(p+1)/2}}{\Gamma(\frac{p+1}{2})}+\gamma \\ +[\log(1+e^{-\xi})+\tfrac{\xi-\gamma}{2}+\tfrac{1/2-g(\xi)}{2\xi}(\xi^2-\gamma^2)]\tfrac{2\pi^{(p+1)/2}}{\Gamma(\frac{p+1}{2})}\end{aligned} \tag{7}$$

The expectation of this upper bound is much easier to compute. Specifically, we need to tackle $\mathbb{E}_{q(\mathbf{A})}[\|\sum_{j=1}^{i-1}\tilde{\mathbf{a}}_j\|_2^2]$, which can be computed as

$$\begin{aligned}&\mathbb{E}_{q(\mathbf{A})}[\|\textstyle\sum_{j=1}^{i-1}\tilde{\mathbf{a}}_j\|_2^2] \\ &= \mathbb{E}_{q(\mathbf{A})}[\textstyle\sum_{j=1}^{i-1}\tilde{\mathbf{a}}_j^\top\tilde{\mathbf{a}}_j+\sum_{j=1}^{i-1}\sum_{k\neq j}^{i-1}\tilde{\mathbf{a}}_j^\top\tilde{\mathbf{a}}_k] \\ &= \sum_{j=1}^{i-1}\mathrm{tr}(\mathbb{E}_{q(\tilde{\mathbf{a}}_j)}[\tilde{\mathbf{a}}_j\tilde{\mathbf{a}}_j^\top])+\sum_{j=1}^{i-1}\sum_{k\neq j}^{i-1}\mathbb{E}_{q(\tilde{\mathbf{a}}_j)}[\tilde{\mathbf{a}}_j]^\top\mathbb{E}_{q(\tilde{\mathbf{a}}_k)}[\tilde{\mathbf{a}}_k] \\ &= \sum_{j=1}^{i-1}\mathrm{tr}(\mathrm{cov}(\tilde{\mathbf{a}}_j))+\sum_{j=1}^{i-1}\sum_{k=1}^{i-1}\mathbb{E}_{q(\tilde{\mathbf{a}}_j)}[\tilde{\mathbf{a}}_j]^\top\mathbb{E}_{q(\tilde{\mathbf{a}}_k)}[\tilde{\mathbf{a}}_k]\end{aligned} \tag{8}$$

where $\mathbb{E}_{q(\tilde{\mathbf{a}}_j)}[\tilde{\mathbf{a}}_j] = A_p(\hat{\kappa})\hat{\mathbf{a}}_j$, $\mathrm{cov}(\tilde{\mathbf{a}}_j) = \frac{h(\hat{\kappa})}{\hat{\kappa}}\mathbf{I}+(1-2\frac{\nu+1}{\hat{\kappa}}h(\hat{\kappa})-h^2(\hat{\kappa}))\hat{\mathbf{a}}_j\hat{\mathbf{a}}_j^\top$, $h(\hat{\kappa}) = \frac{I_{\nu+1}(\hat{\kappa})}{I_\nu(\hat{\kappa})}$, $A_p(\hat{\kappa}) = \frac{I_{p/2}(\hat{\kappa})}{I_{p/2-1}(\hat{\kappa})}$ and $\nu = p/2-1$.

MCMC Sampling One potential drawback of the variational inference approach is that large approximation error can be incurred if the variational distribution is far from the true posterior. In this section, we study an alternative approximation inference method — Markov chain Monte Carlo (MCMC) (Gilks, 2005), which draws samples from the *exact* posterior distribution and uses the samples to represent the posterior. Specifically we choose the Metropolis-Hastings (MH) algorithm (Gilks, 2005) which generates samples from an adaptive proposal distribution, computes acceptance probabilities based on the unnormalized true posterior and uses the acceptance probabilities to decide whether a sample should be accepted or rejected. MH eventually converges to a stationary distribution where the generated samples represent the true posterior. The downside of MCMC is that it could take a long time to converge, which is usually computationally less efficient than variational inference (Hoffman et al., 2013). For the directional variables $\{\tilde{\mathbf{a}}_i\}_{i=1}^K$ and magnitude variables $\{g_i\}_{i=1}^K$, we define the proposal distributions to be von Mises-Fisher and normal distribution

$$\begin{aligned}q(\tilde{\mathbf{a}}_i^{(t+1)}|\tilde{\mathbf{a}}_i^{(t)}) = C_p(\hat{\kappa})\exp(\hat{\kappa}\tilde{\mathbf{a}}_i^{(t+1)}\cdot\tilde{\mathbf{a}}_i^{(t)}) \\ q(g_i^{(t+1)}|g_i^{(t)}) = \tfrac{1}{\sigma\sqrt{2\pi}}\exp\{-\tfrac{(g_i^{(t+1)}-g_i^{(t)})^2}{2\sigma^2}\}\end{aligned} \tag{9}$$

Under the MH algorithm, the MABN prior facilitates better efficiency compared with the DPP prior. In each iteration, the MABN prior needs to be evaluated, whose complexity is quadratic in the number of components K whereas evaluating the DPP has a cubic complexity in K.

2.2. Diversity-Promoting Posterior Regularization

In practice, one may desire to achieve more than one diversity-promoting effects in LVMs. For example, the mutual angular regularizer (Xie et al., 2015) aims to encourage the pairwise angles between components to have not only large mean, but also small variance such that the components are uniformly "different" from each other and evenly spread out to different directions in the space. It would be extremely difficult, if ever possible, to define a proper prior that can accommodate all desired effects. For instance, the MABN priors defined above can encourage the mutual angles to have large mean, but are unable to promote small variance. To overcome such inflexibility of the prior control method, we resort to a *posterior regularization* approach (Zhu et al., 2014). Instead of designing a Bayesian prior to encode the diversification desideratum and indirectly influencing the posterior, posterior regularization directly imposes a control over the post-data distributions to achieve certain goals. Giving prior $\pi(\mathbf{A})$ and data likelihood $p(\mathcal{D}|\mathbf{A})$, computing the posterior $p(\mathbf{A}|\mathcal{D})$ is equivalent to solving the following optimization problem (Zhu et al., 2014)

$$\sup_{q(\mathbf{A})}\mathbb{E}_{q(\mathbf{A})}[\log p(\mathcal{D}|\mathbf{A})\pi(\mathbf{A})]-\mathbb{E}_{q(\mathbf{A})}[\log q(\mathbf{A})] \tag{10}$$

where $q(\mathbf{A})$ is any valid probability distribution. The basic idea of posterior regularization is to impose a certain regularizer $\mathcal{R}(q(\mathbf{A}))$ over $q(\mathbf{A})$ to incorporate prior knowledge and structural bias (Zhu et al., 2014) and solve the following regularized problem

$$\begin{aligned}\sup_{q(\mathbf{A})}\mathbb{E}_{q(\mathbf{A})}[\log p(\mathcal{D}|\mathbf{A})\pi(\mathbf{A})]-\mathbb{E}_{q(\mathbf{A})}[\log q(\mathbf{A})] \\ +\lambda\mathcal{R}(q(\mathbf{A}))\end{aligned} \tag{11}$$

where λ is a tradeoff parameter. Through properly designing $\mathcal{R}(q(\mathbf{A}))$, many diversity-promoting effects can be flexibly incorporated. Here we present a specific example while noting that many other choices are applicable. Gaining insight from (Xie et al., 2015), we define $\mathcal{R}(q(\mathbf{A}))$ as

$$\begin{aligned}\Omega(\{\mathbb{E}_{q(\mathbf{a}_i)}[\mathbf{a}_i]\}_{i=1}^K) = \tfrac{1}{K(K-1)}\sum_{i=1}^K\sum_{j\neq i}^K\theta_{ij}-\gamma\tfrac{1}{K(K-1)} \\ \sum_{i=1}^K\sum_{j\neq i}^K(\theta_{ij}-\tfrac{1}{K(K-1)}\sum_{p=1}^K\sum_{q\neq p}^K\theta_{pq})^2\end{aligned} \tag{12}$$

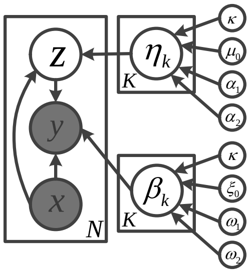

Figure 3. Bayesian Mixture of Experts with Mutual Angular Prior

where $\theta_{ij} = \arccos(\frac{|\mathbb{E}[\mathbf{a}_i]\cdot\mathbb{E}[\mathbf{a}_j]|}{\|\mathbb{E}[\mathbf{a}_i]\|_2\|\mathbb{E}[\mathbf{a}_j]\|_2})$ is the non-obtuse angle measuring the dissimilarity between $\mathbb{E}[\mathbf{a}_i]$ and $\mathbb{E}[\mathbf{a}_j]$, and the regularizer is defined as the mean of pairwise angles minus their variance. The intuition behind this regularizer is: if the mean of angles is larger (indicating these vectors are more different from each other on the whole) and the variance of the angles is smaller (indicating these vectors evenly spread out to different directions), then these vectors are more diverse. Note that it is very difficult to design priors to simultaneously achieve these two effects.

While posterior regularization is more flexible, it lacks some strengths possessed by the prior control method. First, prior control is a more natural way of incorporating prior knowledge, with solid theoretical foundation. Second, prior control can facilitate sampling based algorithms that are not applicable in posterior regularization. In sum, the two approaches have complementary advantages and should be chosen according to specific problem context.

2.3. "Diversifying" Bayesian Mixture of Experts Model

In this section, we apply the two approaches developed above to "diversify" the Bayesian mixture of experts model (BMEM) (Waterhouse et al., 1996).

2.3.1. BMEM with Mutual Angular Prior

Mixture of experts model (MEM) (Jordan & Jacobs, 1994) has been widely used for machine learning tasks where the distribution of input data is so complicated that a single model ("expert") cannot be effective for all the data. MEM assumes that the input data is inherently belonging to multiple latent groups and one single "expert" is allocated to each group to handle the data therein. Here we consider a classification task whose goal is to learn binary linear classifiers given the training data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ where $\mathbf{x}$ is input feature vector and $y_i \in \{1, 0\}$ is class label. We assume there are K latent experts where each expert is a classifier with coefficient vector $\boldsymbol{\beta}$. Given a test sample $\mathbf{x}$, it first goes through a gate function that decides which expert is best suitable to classify this sample and the decision is made in a probabilistic way. A discrete variable z is utilized to indicate the selected expert and the probability that $z = k$ (assigning sample $\mathbf{x}$ to expert k) is $\frac{\exp(\boldsymbol{\eta}_k^\top \mathbf{x})}{\sum_{j=1}^K \exp(\boldsymbol{\eta}_j^\top \mathbf{x})}$ where $\boldsymbol{\eta}_k$ is a coefficient vector characterizing the selection of expert k. Given the selected expert, the sample is classified using the coefficient vector $\boldsymbol{\beta}$ corresponding to that expert. As described in Figure 3, the generative process of $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ is as follows

- For $i = 1, \cdots, N$
  - Draw $z_i \sim \text{Multi}(\boldsymbol{\zeta})$, $\zeta_k = \frac{\exp(\boldsymbol{\eta}_k^\top \mathbf{x}_i)}{\sum_{j=1}^K \exp(\boldsymbol{\eta}_j^\top \mathbf{x}_i)}$
  - Draw $y_i \sim \text{Bernoulli}(\frac{1}{1+\exp(-\boldsymbol{\beta}_{z_i}^\top \mathbf{x}_i)})$

As of now, the model parameters $\mathbf{B} = \{\boldsymbol{\beta}_k\}_{k=1}^K$ and $\mathbf{H} = \{\boldsymbol{\eta}_k\}_{k=1}^K$ are deterministic variables. Next we place a prior over them to enable Bayesian learning (Waterhouse et al., 1996) and desire this prior to be able to promote diversity among the experts to retain the advantages of "diversifying" LVMs as stated before. The mutual angular Bayesian network prior can be applied to achieve this goal

$$\begin{aligned} p(\mathbf{B}) = {} & C_p(\kappa)\exp(\kappa\mu_0^\top \tilde{\boldsymbol{\beta}}_1) \textstyle\prod_{i=2}^K C_p(\kappa) \\ & \exp(\kappa(-\frac{\sum_{j=1}^{i-1}\tilde{\boldsymbol{\beta}}_j}{||\sum_{j=1}^{i-1}\tilde{\boldsymbol{\beta}}_j||_2})^\top \tilde{\boldsymbol{\beta}}_i) \textstyle\prod_{i=1}^K \frac{\alpha_2^{\alpha_1} g_i^{\alpha_1 - 1} e^{-g_i \alpha_2}}{\Gamma(\alpha_1)} \end{aligned} \quad (13)$$

$$\begin{aligned} p(\mathbf{H}) = {} & C_p(\kappa)\exp(\kappa\xi_0^\top \tilde{\boldsymbol{\eta}}_1) \textstyle\prod_{i=2}^K C_p(\kappa) \\ & \exp(\kappa(-\frac{\sum_{j=1}^{i-1}\tilde{\boldsymbol{\eta}}_j}{||\sum_{j=1}^{i-1}\tilde{\boldsymbol{\eta}}_j||_2})^\top \tilde{\boldsymbol{\eta}}_i) \textstyle\prod_{i=1}^K \frac{\omega_2^{\omega_1} h_i^{\omega_1 - 1} e^{-h_i \omega_2}}{\Gamma(\omega_1)} \end{aligned} \quad (14)$$

where $\boldsymbol{\beta}_k = g_k \tilde{\boldsymbol{\beta}}_k$ and $\boldsymbol{\eta}_k = h_k \tilde{\boldsymbol{\eta}}_k$.

2.3.2. BMEM with Mutual Angular Posterior Regularization

As an alternative approach, the diversity in BMEM can be imposed by placing the mutual angular regularizer (Eq.(12)) over the post-data posteriors (Zhu et al., 2014). Here we instantiate the general diversity-promoting posterior regularization defined in Eq.(11) to BMEM, by specifying the following parametrization. The latent variables in BMEM include $\mathbf{B}$, $\mathbf{H}$ and $\mathbf{z} = \{z_i\}_{i=1}^N$ and the post-data distribution over them is defined as $q(\mathbf{B}, \mathbf{H}, \mathbf{z}) = q(\mathbf{B})q(\mathbf{H})q(\mathbf{z})$. For computational tractability, we define $q(\mathbf{B})$ and $q(\mathbf{H})$ to be: $q(\mathbf{B}) = \prod_{k=1}^K q(\tilde{\boldsymbol{\beta}}_k)q(g_k)$ and $q(\mathbf{H}) = \prod_{k=1}^K q(\tilde{\boldsymbol{\eta}}_k)q(h_k)$ where $q(\tilde{\boldsymbol{\beta}}_k)$, $q(\tilde{\boldsymbol{\eta}}_k)$ are von Mises-Fisher distributions and $q(g_k)$, $q(h_k)$ are gamma distributions, and define $q(\mathbf{z})$ to be multinomial distributions: $q(\mathbf{z}) = \prod_{i=1}^N q(z_i|\boldsymbol{\phi}_i)$ where $\boldsymbol{\phi}_i$ is a multinomial vector. The priors over $\mathbf{B}$ and $\mathbf{H}$ are specified to be: $\pi(\mathbf{B}) = \prod_{k=1}^K p(\tilde{\boldsymbol{\beta}}_k)p(g_k)$ and $\pi(\mathbf{H}) = \prod_{k=1}^K p(\tilde{\boldsymbol{\eta}}_k)p(h_k)$ where $p(\tilde{\boldsymbol{\beta}}_k)$, $p(\tilde{\boldsymbol{\eta}}_k)$ are vMF distributions and $p(g_k)$, $p(h_k)$ are gamma distributions. Under such parametrization, we solve the following diversity-promoting posterior regularization problem

$$\begin{aligned}\sup_{q(\mathbf{B},\mathbf{H},\mathbf{z})}\ & \mathbb{E}_{q(\mathbf{B},\mathbf{H},\mathbf{z})}[\log p(\{y_i\}_{i=1}^{N},\mathbf{z}|\mathbf{B},\mathbf{H})\pi(\mathbf{B},\mathbf{H})]\\ & -\mathbb{E}_{q(\mathbf{B},\mathbf{H},\mathbf{z})}[\log q(\mathbf{B},\mathbf{H},\mathbf{z})]\\ & +\lambda_1\Omega(\{\mathbb{E}_{q(\tilde{\boldsymbol{\beta}}_k)}[\tilde{\boldsymbol{\beta}}_k]\}_{k=1}^{K})\\ & +\lambda_2\Omega(\{\mathbb{E}_{q(\tilde{\boldsymbol{\eta}}_k)}[\tilde{\boldsymbol{\eta}}_k]\}_{k=1}^{K})\end{aligned} \qquad (15)$$

Note that other parametrizations are also valid, such as placing Gaussian priors over $\mathbf{B}$ and $\mathbf{H}$ and setting $q(\mathbf{B})$, $q(\mathbf{H})$ to be Gaussian.

3. Experiments

Using Bayesian mixture of experts model as an instance, we conducted experiments to verify the effectiveness and efficiency of the two proposed approaches.

Datasets We used two binary-classification datasets. The first one is the Adult-9 (Platt et al., 1999) dataset, which has ~33K training instances and ~16K testing instances. The feature dimension is 123. The other dataset is SUN-Building compiled from the SUN (Xiao et al., 2010) dataset, which contains ~6K building images and 7K non-building images randomly sampled from other categories, where 70% of images are used for training and the rest for testing. We use SIFT (Lowe, 1999) based bag-of-words to represent the images with a dimensionality of 1000.

Experimental Settings To understand the effects of diversification in Bayesian learning, we compare the following methods: (1) mixture of experts model (MEM) with L2 regularization (MEM-L2) where the L2 regularizer is imposed over "experts" independently; (2) MEM with mutual angular regularization (Xie et al., 2015) (MEM-MAR) where the "experts" are encouraged to be diverse; (3) Bayesian MEM with a Gaussian prior (BMEM-G) where the "experts" are independently drawn from a Gaussian prior; (4) BMEM with mutual angular Bayesian network priors (type I or II) where the MABN favors diverse "experts" (BMEM-MABN-I, BMEM-MABN-II); BMEM-MABN-I is inferred with MCMC sampling and BMEM-MABN-II is inferred with variational inference; (5) BMEM with posterior regularization (BMEM-PR).

The key parameters involved in the above methods are: (1) the regularization parameter λ in MEM-L2, MEM-MAR, BMEM-PR; (2) the concentration parameter κ in the mutual angular priors in BMEM-MABN-(I,II); (3) the concentration parameter $\hat{\kappa}$ in the variational distribution in BMEM-MABN-II. All parameters were tuned using 5-fold cross validation. Besides internal comparison, we also compared with 5 baseline methods, which are among the most widely used classification approaches that achieve the state of the art performance. They are: (1) kernel support vector machine (KSVM) (Burges, 1998); (2) random forest (RF) (Breiman, 2001); (3) deep neural network (DNN) (Hinton & Salakhutdinov, 2006); (4) Infinite SVM (ISVM) (Zhu et al., 2011); (5) BMEM with DPP prior (BMEM-DPP) (Kulesza & Taskar, 2012), in which a Metropolis-Hastings sampling algorithm was adopted[1]. The kernels in KSVM and BMEM-DPP are both radial basis function kernel. Parameters of the baselines were tuned using 5-fold cross validation.

K	5	10	20	30
MEM-L2	82.6	83.8	84.3	84.7
MEM-MAR	85.3	86.4	86.6	87.1
BMEM-G	83.4	84.2	84.9	84.9
BMEM-MABN-I	**87.1**	**88.3**	88.6	**88.9**
BMEM-MABN-II	86.4	87.8	88.1	88.4
BMEM-PR	86.2	87.9	**88.7**	88.1

Table 1. Classification accuracy (%) on Adult-9 dataset

K	5	10	20	30
MEM-L2	76.2	78.8	79.4	79.7
MEM-MAR	81.3	82.1	82.7	82.3
BMEM-G	76.5	78.6	80.2	80.4
BMEM-MABN-I	**82.1**	83.6	**85.3**	**85.2**
BMEM-MABN-II	80.9	82.8	84.9	84.1
BMEM-PR	81.7	**84.1**	83.8	84.9

Table 2. Classification accuracy (%) on SUN-Building dataset

Results Table 1 and 2 show the classification accuracy under different number of "experts" on the Adult-9 and SUN-Building dataset respectively. From these two tables, we observe that: (1) diversification can greatly improve the performance of Bayesian MEM, which can be seen from the comparison between diversified BMEM methods and their non-diversified counterparts, such as BMEM-MABN-(I,II) versus BMEM-G, and BMEM-PR versus BMEM-G. (2) Bayesian learning achieves better performance than point estimation, which is manifested by comparing BMEM-G with MEM-L2, and BMEM-MABN-(I,II)/BMEM-PR with MEM-MAR. (3) BMEM-MAR-I works better than BMEM-MABN-II and BMEM-PR. The reason is that BMEM-MAR-I inferred with MCMC draws samples from the *exact* posterior while BMEM-MABN-II and BMEM-PR inferred with variational inference seek an *approximation* of the posterior.

Recall that the goals of promoting diversity in LVMs are to reduce model size without sacrificing modeling power and effectively capture infrequent patterns. Here we empirically verify whether these two goals can be achieved. Regarding the first goal, we compare diversified BMEM methods BMEM-MABN-(I,II)/BMEM-PR with non-diversified counterpart BMEM-G and check whether

[1] Variational inference and Gibbs sampling (Affandi et al., 2013) are both not applicable.

Category ID	C18	C17	C12	C14	C22	C34	C23	C32	C16
Num. of Docs	5281	4125	1194	741	611	483	262	208	192
BMEM-G Accuracy (%)	87.3	88.5	75.7	70.1	71.6	64.2	55.9	57.4	51.3
BMEM-MABN-I Accuracy (%)	88.1	86.9	74.7	72.2	70.5	67.3	68.9	70.1	65.5
Relative Improvement (%)	1.0	-1.8	-1.3	2.9	-1.6	4.6	18.9	18.1	21.7

Table 3. Accuracy on 9 subcategories of the CCAT category in the RCV1.Binary dataset

	Adult-9	SUN-Building
KSVM	85.2	84.2
RF	87.7	85.1
DNN	87.1	84.8
ISVM	85.8	82.3
BMEM-DPP	86.5	84.5
BMEM-MABN-I	**88.9**	**85.3**
BMEM-MABN-II	88.4	84.9
BMEM-PR	88.7	84.9

Table 4. Classification accuracy (%) on two datasets

	Adult-9	SUN-Building
BMEM-DPP	8.2	11.7
BMEM-MABN-I	7.5	10.5
BMEM-MABN-II	2.9	4.1
BMEM-PR	3.3	4.9

Table 5. Training time (hours) of different methods with $K = 30$

diversified methods with a small number of components K which entails low computational complexity can achieve performance as good as non-diversified methods with large K. It can be observed that BMEM-MABN-(I,II)/BMEM-PR with a small K can achieve accuracy that is comparable to or even better than BMEM-G with a large K. For example, on the Adult-9 dataset (Table 1), with 5 experts BMEM-MABN-I is able to achieve an accuracy of 87.1%, which cannot be achieved by BMEM-G with even 30 experts. This corroborates the effectiveness of diversification in reducing model size (hence computational complexity) without compromising performance.

To verify the second goal – capturing infrequent patterns, from the RCV1 (Lewis et al., 2004) dataset we pick up a subset of documents (for binary classification) such that the popularity of categories (patterns) is in power-law distribution. Specifically, we choose documents from 9 subcategories (the 1st row of Table 3) of the CCAT category as the positive instances, and randomly sample 15K documents from non-CCAT categories as negative instances. The 2nd row shows the number of documents in each of the 9 categories. The distribution of document frequency is in a power-law fashion, where frequent categories (such as C18 and C17) have a lot of documents while infrequent categories (such as C32 and C16) have a small amount of documents. The 3rd and 4th row show the accuracy achieved by BMEM-G and BMEM-MABN-I on each category. The 5th row shows the relative improvement of BMEM-MABN-I over BMEM-G, which is defined as $\frac{A_{bmem\_mabn}-A_{bmem\_g}}{A_{bmem\_g}}$, where A_{bmem_mabn} and A_{bmem_g} denote the accuracy achieved by BMEM-MABN-I and BMEM-G respectively. While achieving accuracy comparable to BMEM-G over the frequent categories, BMEM-MABN-I obtains much better performance on infrequent categories. For example, the relative improvements on infrequent categories C32 and C16 are 18.1% and 21.7%. This demonstrates that BMEM-MABN-I can effectively capture the infrequent patterns.

Table 4 presents the comparison with the state of the art classification methods. As one can see, our method achieves the best performances on both datasets. In particular, BMEM-MAR-(I,II) work better than BMEM-DPP, demonstrating the proposed mutual angular priors possess ability that is better than or comparable to the DPP prior in inducing diversity.

Table 5 compares the time (hours) taken by each method to achieve convergence, with K set to 30. BMEM-MABN-II inferred with variational inference (VI) is more efficient than BMEM-MABN-I inferred with MCMC sampling, due to the higher efficiency of VI than MCMC. BMEM-PR is solved with an optimization algorithm which is more efficient than the sampling algorithm in BMEM-MABN-I. BMEM-MABN-II and BMEM-PR are more efficient than BMEM-DPP where DPP inhibits the adoption of VI.

4. Conclusions

We study how to promote diversity in Bayesian latent variable models, for the purpose of better capturing infrequent patterns and reducing model size without compromising modeling power. We define mutual angular Bayesian network (MABN) priors entailing bias towards components with larger mutual angles and investigate a posterior regularization approach which directly applies regularizers over the post-data distributions to promote diversity. Approximate algorithms are developed for posterior inference under the MABN priors. With Bayesian mixture of experts model as a study case, empirical experiments demonstrate the effectiveness and efficiency of our methods.

Acknowledgements

P.X and E.X are supported by ONR N00014140684, DARPA XDATA FA8701220324, NSF IIS1447676. J.Z is supported by National NSF of China (Nos. 61322308, 61332007).

References

Affandi, Raja Hafiz, Fox, Emily, and Taskar, Ben. Approximate inference in continuous determinantal processes. In *Advances in Neural Information Processing Systems*, pp. 1430–1438, 2013.

Bishop, Christopher M. Latent variable models. In *Learning in Graphical Models*, pp. 371–403. Springer, 1998.

Bishop, Christopher M and Tipping, Michael E. Bayesian regression and classification. *Nato Science Series sub Series III Computer And Systems Sciences*, 190:267–288, 2003.

Blei, David and Lafferty, John. Correlated topic models. In *Advances in Neural Information Processing Systems*, volume 18, pp. 147. MIT, 2006.

Blei, David M. Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 2014.

Blei, David M, Ng, Andrew Y, and Jordan, Michael I. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003.

Bouchard, Guillaume. Efficient bounds for the softmax function, applications to inference in hybrid models. 2007.

Breiman, Leo. Random forests. *Machine Learning*, 45(1): 5–32, 2001.

Burges, Christopher JC. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

Gilks, Walter R. *Markov chain Monte Carlo*. Wiley Online Library, 2005.

Hinton, Geoffrey E and Salakhutdinov, Ruslan R. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

Hoffman, Matthew D, Blei, David M, Wang, Chong, and Paisley, John. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Jaakkola, T and Jordan, Michael I. A variational approach to bayesian logistic regression models and their extensions. In *Sixth International Workshop on Artificial Intelligence and Statistics*, 1997.

Jordan, Michael I and Jacobs, Robert A. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

Kindermann, Ross, Snell, James Laurie, et al. *Markov random fields and their applications*, volume 1. American Mathematical Society Providence, 1980.

Knott, Martin and Bartholomew, David J. *Latent variable models and factor analysis*. Number 7. Edward Arnold, 1999.

Koller, Daphne and Friedman, Nir. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Kulesza, Alex and Taskar, Ben. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 2012.

Lewis, David D, Yang, Yiming, Rose, Tony G, and Li, Fan. Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, 5:361–397, 2004.

Lowe, David G. Object recognition from local scale-invariant features. In *The Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pp. 1150–1157. Ieee, 1999.

Malkin, Jonathan and Bilmes, Jeff. Ratio semi-definite classifiers. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4113–4116. IEEE, 2008.

Mardia, Kanti V and Jupp, Peter E. *Directional statistics*, volume 494. John Wiley & Sons, 2009.

Neal, Radford M. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Platt, John et al. Fast training of support vector machines using sequential minimal optimization. *Advances in Kernel Methods Support Vector Learning*, 3, 1999.

Wainwright, Martin J and Jordan, Michael I. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2): 1–305, 2008.

Wang, Chong and Blei, David M. Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14(1):1005–1031, 2013.

Wang, Yi, Zhao, Xuemin, Sun, Zhenlong, Yan, Hao, Wang, Lifeng, Jin, Zhihui, Wang, Liubin, Gao, Yang, Law, Ching, and Zeng, Jia. Peacock: Learning long-tail topic features for industrial applications. *ACM Transactions on Intelligent Systems and Technology*, 2014.

Wasserman, Larry. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

Waterhouse, Steve, MacKay, David, Robinson, Tony, et al. Bayesian methods for mixtures of experts. *Advances in Neural Information Processing Systems*, pp. 351–357, 1996.

Xiao, Jianxiong, Hays, James, Ehinger, Krista, Oliva, Aude, Torralba, Antonio, et al. Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492. IEEE, 2010.

Xie, Pengtao. Learning compact and effective distance metrics with diversity regularization. In *European Conference on Machine Learning*, 2015.

Xie, Pengtao, Deng, Yuntian, and Xing, Eric P. Diversifying restricted boltzmann machine for document modeling. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2015.

Zhu, Jun, Chen, Ning, and Xing, Eric P. Infinite svm: a dirichlet process mixture of large-margin kernel machines. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 617–624, 2011.

Zhu, Jun, Chen, Ning, and Xing, Eric P. Bayesian inference with posterior regularization and applications to infinite latent svms. *Journal of Machine Learning Research*, 15 (1):1799–1847, 2014.

Zou, James Y. and Adams, Ryan P. Priors for diversity in generative latent variable models. In *Advances in Neural Information Processing Systems*, 2012.