Bayesian Max-margin Multi-Task Learning with Data Augmentation

[†]**Chengtao Li** CTLI.CS@HOTMAIL.COM
[‡]**Jun Zhu** DCSZJ@MAIL.TSINGHUA.EDU.CN
[‡]**Jianfei Chen** CHENJF10@MAILS.TSINGHUA.EDU.CN
[†]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China
[‡]Dept. of Comp. Sci. & Tech, TNList Lab, State Key Lab of Intell. Tech & Sys, Tsinghua University, Beijing, China

Abstract

Both max-margin and Bayesian methods have been extensively studied in multi-task learning, but have rarely been considered together. We present Bayesian max-margin multi-task learning, which conjoins the two schools of methods, thus allowing the discriminative max-margin methods to enjoy the great flexibility of Bayesian methods on incorporating rich prior information as well as performing nonparametric Bayesian feature learning with the latent dimensionality resolved from data. We develop Gibbs sampling algorithms by exploring data augmentation to deal with the non-smooth hinge loss. For nonparametric models, our algorithms do not need to make mean-field assumptions or truncated approximation. Empirical results demonstrate superior performance than competitors in both multi-task classification and regression.

1. Introduction

Multi-task learning (MTL) (Caruana, 1997; Thrun & O'Sullivan, 1996) has been widely studied in computer vision (Zhang et al., 2012), text classification (Zhu et al., 2013), and bioinformatics (Widmer & Rätsch, 2012). MTL is particularly good for the scenarios where some tasks are under sampled. The primary belief of MTL is that solving multiple potentially correlated tasks together could improve the performance of some (or all) of these tasks by sharing statistic strength. According to the strategies on sharing statistics, existing methods can be grouped into two categories. The first category consists of approaches that aim to discover the relationships between tasks (Bakker & Heskes, 2003; Jacob et al., 2008; Xue et al., 2007; Yu et al., 2007; Hariharan et al., 2010), while the other one consists of the approaches that aim to mine the related features or find a common feature structure shared by all tasks (Argyriou et al., 2006; Chen et al., 2009; Rai & Daume, 2010; Obozinski et al., 2010). Recent attempts have been made to simultaneously estimate task correlations and feature correlations in one unified learning method, under either a Bayesian or a max-margin framework. Bayesian methods (Archambeau et al., 2011; Yang et al., 2013) employ hierarchical structures to model multi-task data and try to extract both types of correlations by setting proper priors. Though enjoying great flexibility for incorporating latent variables and performing nonparametric Bayesian inference, the generative nature could make these Bayesian methods less than sufficient in predictive learning. On the other hand, the max-margin MTL methods (Zhang & Schneider, 2010) could also learn task correlations and feature correlations simultaneously by using a proper regularization over model parameters, under a regularized loss minimization framework. Though max-margin methods enjoy the strong discriminative ability, they are usually lack of the flexibility of Bayesian methods.

One recent work that attempts to bring max-margin learning and Bayesian multi-task learning together is the multi-task infinite latent SVM (MT-iLSVM) (Zhu et al., 2014b), which learns a common projection matrix to share statistics among multiple tasks. By performing max-margin learning, it could achieve promising prediction results. It also applies nonparametric Bayesian methods to automatically resolve the dimensionality of the projection matrix. However, MT-iLSVM solely focuses on learning latent features, while not considering task correlations. Furthermore, for computational tractability, MT-iLSVM adopts variational approximation methods with truncated mean-field assumptions, which could be too strict to be realistic in practice.

This paper presents a systematical study of Bayesian max-margin multi-task learning (BM-MTL) for both classification and regression. First, we present a generic framework of performing Bayesian max-margin MTL, which can

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

use structured priors, e.g., the matrix normal prior (Archambeau et al., 2011), to jointly estimate task correlations and feature correlations. Second, we extend the basic BM-MTL framework and present a nonparametric Bayesian max-margin MTL method (NPBM-MTL), which can learn latent feature representations and estimate task correlations, with latent dimensionality automatically resolved from data. Finally, for all the Bayesian max-margin MTL methods, we develop simple Gibbs sampling algorithms by exploring data augmentation techniques (Tanner & Wong, 1987; Polson & Scott, 2011; Zhu et al., 2014a). Unlike the truncated variational mean-field methods of MTiLSVM, our algorithms for the nonparametric Bayesian methods do not make any restrictive assumptions and are truncation free, thus allowing for inference of the true posterior distributions. We empirically study the effectiveness of our methods and make comparisons with other state-of-the-art models. Our results on several real data sets demonstrate superior performance than various competitors in both multi-task classification and regression tasks.

2. Background

We briefly overview Multi-Task SVM and Bayesian MTL approaches, on which our methods are based.

2.1. Multi-Task SVM and Extensions

Let L be the number of tasks. We denote the multi-task training set by $\mathcal{D} = (\mathbf{X}, \mathbf{Y}) = \{\{(\mathbf{x}_{il}, y_{il})\}_{i=1}^{N_l}\}_{l=1}^{L}$, where N_l is the number of instances in task l. Each data instance is a pair $(\mathbf{x}_{il}, y_{il})$ with $\mathbf{x}_{il} \in \mathbb{R}^D$ being an input feature vector and y_{il} being a response variable. Without loss of generality, we consider binary classification, where y_{il} equals to $+1$ if the label of instance i in task l is positive and -1 otherwise. For a linear MTL model, we characterize each learning task l by a parameter vector $\boldsymbol{\eta}_l \in \mathbb{R}^D$. Let $\boldsymbol{\eta}$ denote the $D \times L$ matrix formed by concatenating $\boldsymbol{\eta}_l$'s of all the tasks. The prediction of the i-th sample in the l-th task is given by the sign rule $\hat{y}_{il} = \text{sgn}(\boldsymbol{\eta}_l^\top \mathbf{x}_{il})$, where $\text{sgn}(x)$ is $+1$ if $x \geq 0$ and -1 otherwise. Multi-task SVM (MTSVM) employs hinge loss as its loss measure for each instance, i.e., for the i-th sample in task l, the loss is calculated as $\max(0, \zeta_{il})$, where $\zeta_{il} = t - y_{il}\boldsymbol{\eta}_l^\top \mathbf{x}_{il}$ with t specifying the panelty of making a wrong prediction. The objective function of MTSVM is then defined as

$$\min_{\boldsymbol{\eta}} \frac{1}{2} \text{Reg}(\boldsymbol{\eta}) + 2C \sum_{l=1}^{L} \sum_{i=1}^{N_l} \big[\max(0, \zeta_{il})\big], \qquad (1)$$

where $\text{Reg}(\boldsymbol{\eta})$ is the regularization term over $\boldsymbol{\eta}$. A naïve choice of the regularization is $\| \text{Vec}(\boldsymbol{\eta})\|^2$, where $\text{Vec}(\boldsymbol{\eta})$ is the vectorization of $\boldsymbol{\eta}$ with dimension $DL \times 1$. With such a regularization the model degenerates to a set of single-task SVMs, one for each task, thus doesn't share statistics among tasks. Many efforts have been done to fit this max-margin framework to multi-task learning problems better and make good use of hidden correlation information. Previous work has focused on designing appropriate regularization terms to capture the correlations among tasks and features. For example, Argyriou et al. (2006) used the regularization of $\text{Tr}\{\boldsymbol{\eta}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{\eta}\}$ to model feature correlations via the *feature correlation matrix* $\boldsymbol{\Omega}$; Hariharan et al. (2010) and Zhang & Yeung (2012) used $\text{Tr}\{\boldsymbol{\eta}\mathbf{R}^{-1}\boldsymbol{\eta}^\top\}$ as regularization to correlate $\boldsymbol{\eta}_l$'s together via the *task correlation matrix* $\mathbf{R}$; and Zhang & Schneider (2010) proposed to use $\text{Tr}\{\boldsymbol{\Omega}^{-1}\boldsymbol{\eta}\mathbf{R}^{-1}\boldsymbol{\eta}^\top\}$ as the regularization term, and added additional parameters related to $|\mathbf{R}|$ and $|\boldsymbol{\Omega}|$ to learn sparse correlations among both tasks and features.

Though powerful on learning discriminative models, all the aforementioned methods perform point estimation to learn a single large-margin model by solving a regularized loss minimization problem. Such a deterministic formulation could limit the model flexibility. For instance, it lacks the flexibility of Bayesian methods on incorporating latent variables, leveraging informative priors, or performing nonparametric inference, as reviewed later.

2.2. Bayesian Multi-task Learning

Due to the great flexibility for incorporating rich priors and performing nonparametric inference, Bayesian methods have been widely used for multi-task learning. For example, Xue et al. (2007) employed a nonparametric hierarchical structure in Bayesian framework to automatically adjust model complexity and learn the shared statistics among tasks; Archambeau et al. (2011) assumed a matrix normal prior over the model parameters with the mean being decomposed into two latent matrices and with the covariance matrices following an inverse Wishart distribution; later Yang et al. (2013) extended such approaches with the covariance matrices of model parameter following a more complex matrix generalized inverse Gaussian (Barndorff-Nielsen et al., 1982) prior. Bayesian models allow one to learn latent structures hiding in the data, and by placing smart prior distributions like matrix normal over model parameters, one would be able to infer the task and feature correlations simply by doing Bayesian inference. However, such Bayesian models have difficulty to incorporate side information like discriminative margin constraints and structural bias, thus may suffer in performance. Finally, under a similar constrained Bayesian inference framework, Koyejo & Ghosh (2013) presented a Bayesian MTL method, where a nuclear norm constraint on the predictive weight matrix was used to force a low rank solution.

3. Bayesian Max-margin Multi-task Learning

The advantages of max-margin learning and Bayesian methods could be integrated by bringing them together and do Bayesian max-margin multi-task learning (BM-MTL).

We now present a BM-MTL method to jointly estimate task correlations and feature correlations.

3.1. The Model

As a Bayesian model, we learn a posterior distribution of the classifier weights, $q(\boldsymbol{\eta})$. We adopt the approach of Gibbs classifiers (McAllester, 2003; Germain et al., 2009) to account for the uncertainty of the models. Specifically, if a random sample $\boldsymbol{\eta}$ is drawn from $q(\boldsymbol{\eta})$, we can make predictions using the same sign rule as in the deterministic MT-SVM, and measure the goodness of the sampled classifier using the hinge-loss as a surrogate for the training error. Gibbs classifiers consider the posterior distribution by taking the expectation and define the expected hinge loss

$$\mathcal{R}\big(q(\boldsymbol{\eta})\big) = \sum_{l=1}^{L}\sum_{i=1}^{N_l} \mathbb{E}_q\big[\max(0, \zeta_{il})\big], \tag{2}$$

which is a good surrogate (in fact, an upper bound) for the expected training error, $t\sum_{l=1}^{L}\sum_{i=1}^{N_l}\mathbb{E}_q[\mathbb{I}(\hat{y}_{il} \neq y_{il})]$. Given a prior distribution over $\boldsymbol{\eta}$, $p_0(\boldsymbol{\eta})$, we define the Bayesian max-margin multi-task learning (BM-MTL) as solving the entropy-regularized loss minimization problem

$$\min_{q(\boldsymbol{\eta})\in\mathcal{P}} \mathcal{L}\big(q(\boldsymbol{\eta})\big) + 2C\cdot\mathcal{R}\big(q(\boldsymbol{\eta})\big), \tag{3}$$

where $\mathcal{L}(q(\boldsymbol{\eta})) = \mathrm{KL}(q(\boldsymbol{\eta})\|p_0(\boldsymbol{\eta}))$ and $\mathcal{P}$ is the probability simplex with an appropriate dimension.

We should note that although the entropy-regularized loss minimization problem looks similar to that of maximum entropy discrimination (MED) (Jaakkola et al., 1999), this loss function derived from a Gibbs classifier has rarely been studied. As we shall see, it will lead to simple Gibbs sampling algorithms by exploring data augmentation. Moreover, in a Bayesian formulation, we have the flexibility to incorporate a likelihood model if necessary. Let $p(\mathcal{D}|\boldsymbol{\gamma})$ be a likelihood parameterized by $\boldsymbol{\gamma}$. We can perform regularized Bayesian inference by solving the augmented problem

$$\min_{q(\boldsymbol{\eta},\boldsymbol{\gamma})\in\mathcal{P}} \mathcal{L}\big(q(\boldsymbol{\eta},\boldsymbol{\gamma})\big) + 2C\cdot\mathcal{R}\big(q(\boldsymbol{\eta})\big), \tag{4}$$

where $\mathcal{L}(q(\boldsymbol{\eta},\boldsymbol{\gamma})) = \mathrm{KL}(q\|p_0(\boldsymbol{\eta},\boldsymbol{\gamma})) - \mathbb{E}_q[\log p(\mathcal{D}|\boldsymbol{\gamma})]$.

Let $\psi(y_{il}|\boldsymbol{\eta}, \mathbf{x}_{il}) = \exp\{-2C\max(0,\zeta_{il})\}$ be the unnormalized likelihood of y_{il} for the i-th sample in l-th task. Then solving problem (3) will result in the posterior distribution $q(\boldsymbol{\eta}|\mathcal{D}) = p_0(\boldsymbol{\eta})\psi(\mathbf{Y}|\boldsymbol{\eta},\mathbf{X})/\Gamma(\mathcal{D})$, where $\psi(\mathbf{Y}|\boldsymbol{\eta},\mathbf{X}) = \prod_{l=1}^{L}\prod_{i=1}^{N_l}\psi(y_{il}|\boldsymbol{\eta},\mathbf{x}_{il})$ and $\Gamma(\mathcal{D})$ is a normalization factor to make q a normalized probability distribution. This update rule is in a similar form as Bayes' rule. Such a transformation from max-margin learning to Bayesian inference results in discriminative models that inherit the flexibility of Bayesian methods. Comparing with standard Bayesian methods, our method has the flexibility of incorporating rich side-information such as max-margin posterior regularization, which is hard to be done within full Bayesian framework.

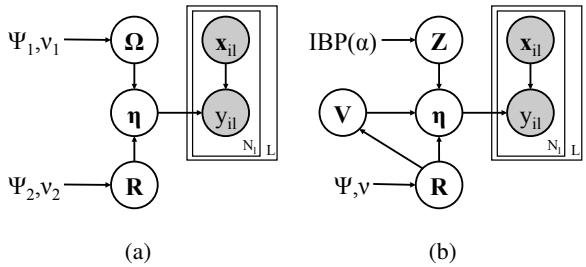

Figure 1. Graphical models for (a) BM-MTL; (b) NPBM-MTL.

The regularization term in MTSVM within this Bayesian formalism could be explained as a prior distribution over $\boldsymbol{\eta}$. The naïve regularization of $\|\mathrm{Vec}(\boldsymbol{\eta})\|^2$ corresponds to a prior of multi-variate normal distribution over $\mathrm{Vec}(\boldsymbol{\eta})$. As we have stated, such a regularization (prior) is ignorant of the fact that $\boldsymbol{\eta}$ is actually a $D\times L$ matrix instead of a single column vector and results in a covariance of $\boldsymbol{\eta}$ of size $DL\times DL$, which is usually prohibitive for modeling and estimation. To capture the structure of $\boldsymbol{\eta}$, *matrix normal prior* could be used (Yang et al., 2013; Archambeau et al., 2011), which assume that the $DL\times DL$ covariance matrix could be decomposed as a Kronecker product $\boldsymbol{\Omega}\otimes\mathbf{R}$, and $\boldsymbol{\eta}$ follows $\mathrm{Vec}(\boldsymbol{\eta}) \sim \mathcal{N}_{D,L}(\mathbf{0}, \boldsymbol{\Omega}\otimes\mathbf{R})$, i.e.,

$$p_0(\boldsymbol{\eta}|\boldsymbol{\Omega},\mathbf{R}) = \frac{\exp\left\{-\frac{1}{2}\mathrm{Tr}(\boldsymbol{\Omega}^{-1}\boldsymbol{\eta}\mathbf{R}^{-1}\boldsymbol{\eta}^\top)\right\}}{(2\pi)^{DL/2}|\mathbf{R}|^{D/2}|\boldsymbol{\Omega}|^{L/2}}. \tag{5}$$

Such a prior has the advantage in that it could use decomposed covariance matrix to model latent correlations in data, where $\boldsymbol{\Omega}$ corresponds to feature correlations and $\mathbf{R}$ corresponds to task correlations respectively, as described in (Zhang & Schneider, 2010).

The Bayesian framework allows us to view correlation matrices as random variables and include proper priors for them. There are many possible choices (Archambeau et al., 2011; Yang et al., 2013). We adopt the conjugate inverse Wishart priors (Mardia et al., 1980), namely, $\boldsymbol{\Omega} \sim \mathcal{IW}(\boldsymbol{\Psi}_1, \nu_1)$ and $\mathbf{R} \sim \mathcal{IW}(\boldsymbol{\Psi}_2, \nu_2)$. Then, the posterior probability becomes

$$\begin{aligned} q(\boldsymbol{\eta},\boldsymbol{\Omega},\mathbf{R}|\mathcal{D}) \propto\ & p_0(\boldsymbol{\Omega}|\boldsymbol{\Psi}_1,\nu_1)p_0(\mathbf{R}|\boldsymbol{\Psi}_2,\nu_2) \\ & \times p_0(\boldsymbol{\eta}|\boldsymbol{\Omega},\mathbf{R})\psi(\mathbf{Y}|\boldsymbol{\eta},\mathbf{X}). \end{aligned} \tag{6}$$

Fig. 1(a) shows the graphical structure of the model.

3.2. Gibbs Sampling with Data Augmentation

Although the expected hinge loss $\mathcal{R}$ is hard to deal with, by using ideas of data augmentation (Tanner & Wong, 1987; Polson & Scott, 2011) we can express the posterior distribution as a marginal of a higher dimensional distribution

with augmented variables and then develop a simple Gibbs sampler. Specifically, the unnormalized likelihood for each label y_{il} can be expressed as

$$\psi(y_{il}|\boldsymbol{\eta}, \mathbf{x}_{il}) = \int_0^\infty \frac{1}{\sqrt{2\pi\lambda_{il}}} \exp\left\{-\frac{(C\zeta_{il} + \lambda_{il})^2}{2\lambda_{il}}\right\} \mathrm{d}\lambda_{il},$$

with the augmented variable $\lambda_{il} \in (0, \infty)$. This result indicates that the posterior distribution $q(\boldsymbol{\eta}, \boldsymbol{\Omega}, \mathbf{R}|\mathcal{D})$ can be expressed as the marginal of a higher dimensional distribution with the augmented variables $\boldsymbol{\lambda} = \{\{\lambda_{il}\}_{i=1}^{N_l}\}_{l=1}^{L}$, where the complete posterior distribution is

$$q(\boldsymbol{\eta}, \boldsymbol{\Omega}, \mathbf{R}, \boldsymbol{\lambda}|\mathcal{D}) \propto p_0(\boldsymbol{\Omega}) p_0(\mathbf{R}) p_0(\boldsymbol{\eta}|\boldsymbol{\Omega}, \mathbf{R}) \psi(\mathbf{Y}, \boldsymbol{\lambda}|\boldsymbol{\eta}, \mathbf{X})$$

with $\psi(\mathbf{y}, \boldsymbol{\lambda}|\boldsymbol{\eta}, \mathbf{X}) = \prod_{l=1}^{L}\prod_{i=1}^{N_l} \psi(y_{il}, \lambda_{il}|\boldsymbol{\eta}, \mathbf{X})$ and $\psi(y_{il}, \lambda_{il}|\boldsymbol{\eta}, \mathbf{X}) = \frac{1}{\sqrt{2\pi\lambda_{il}}} \exp\{-\frac{1}{2\lambda_{il}}(C\zeta_{il} + \lambda_{il})^2\}$. With the data augmentation representation, we can develop a simple Gibbs sampler to infer the complete distribution $q(\boldsymbol{\eta}, \boldsymbol{\Omega}, \mathbf{R}, \boldsymbol{\lambda}|\mathcal{D})$ and thus the target posterior $q(\boldsymbol{\eta}, \boldsymbol{\Omega}, \mathbf{R}|\mathcal{D})$ by dropping $\boldsymbol{\lambda}$. The Gibbs sampler iteratively performs the following steps:

For $\boldsymbol{\eta}$: we have conditional distribution:

$$q(\boldsymbol{\eta}|\boldsymbol{\Omega}, \mathbf{R}, \boldsymbol{\lambda}) \propto p_0(\boldsymbol{\eta}|\boldsymbol{\Omega}, \mathbf{R}) \psi(\mathbf{Y}, \boldsymbol{\lambda}|\boldsymbol{\eta}, \mathbf{X}).$$

Though jointly sampling $\boldsymbol{\eta}$ will lead to a high-dimensional Gaussian, we can effectively sample each $\boldsymbol{\eta}_l$ task-wisely. This leads to sampling from a low dimensional Gaussian. Namely, we have the conditional probabilities

$$\begin{aligned} q(\boldsymbol{\eta}_l|\boldsymbol{\Omega}, \mathbf{R}, \boldsymbol{\lambda}, \boldsymbol{\eta}_{-l}) \propto & \exp\left\{-\frac{1}{2}\mathrm{Tr}\{\boldsymbol{\Omega}^{-1}\boldsymbol{\eta}\mathbf{R}^{-1}\boldsymbol{\eta}^\top\}\right\} \\ & \times \prod_{i=1}^{N_l} \exp\left\{-\frac{(C\zeta_{il} + \lambda_{il})^2}{2\lambda_{il}}\right\} \\ = & \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l), \end{aligned} \tag{7}$$

where the posterior covariance matrix and mean are computed as $\boldsymbol{\Sigma}_l = (\sum_{i=1}^{N_l} C^2 \frac{\mathbf{x}_{il}\mathbf{x}_{il}^\top}{\lambda_{il}} + \mathbf{R}_{ll}^{-1}\boldsymbol{\Omega}^{-1})^{-1}$ and $\boldsymbol{\mu}_l = \boldsymbol{\Sigma}_l(-\frac{1}{2}\sum_{k\neq l}\mathbf{R}_{lk}^{-1}\boldsymbol{\Omega}^{-1}\boldsymbol{\eta}_k + C\sum_{i=1}^{N_l} y_{il}(\frac{Ct}{\lambda_{il}} + 1)\mathbf{x}_{il})$, respectively. We can easily draw a sample from a D-dimensional normal distribution, and the inverse can be robustly done using Cholesky decomposition, an $O(D^3)$ procedure. Thus the sampling could be done efficiently when D is not very large.

For $\boldsymbol{\Omega}$ and $\mathbf{R}$: due to the conjugacy, we have the inverse Wishart conditional distributions:

$$\begin{aligned} q(\boldsymbol{\Omega}|\mathbf{R}, \boldsymbol{\eta}, \boldsymbol{\lambda}) &= \mathcal{IW}\Big(\boldsymbol{\Psi}_1 + \boldsymbol{\eta}\mathbf{R}^{-1}\boldsymbol{\eta}^\top, \nu_1 + L\Big), \\ q(\mathbf{R}|\boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\lambda}) &= \mathcal{IW}\Big(\boldsymbol{\Psi}_2 + \boldsymbol{\eta}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{\eta}, \nu_2 + D\Big). \end{aligned} \tag{8}$$

The sampling procedure involves matrix inversions of sizes $D \times D$ and $L \times L$ which, again, could be done robustly with Cholesky decomposition. Thus the inversion would be efficient when D and L are not very large.

For $\boldsymbol{\lambda}$: due to the conditional independence, we can sample each λ_{il} seperately from a generalized inverse Gaussian distribution (Devroye, 1986)

$$\begin{aligned} q(\lambda_{il}|\boldsymbol{\eta}, \boldsymbol{\Omega}, \mathbf{R}) &\propto \frac{1}{\sqrt{2\pi\lambda_{il}}} \exp\left\{-\frac{1}{2\lambda_{il}}(C\zeta_{il} + \lambda_{il})^2\right\} \\ &= \mathcal{GIG}\left(\lambda_{il}; \frac{1}{2}, 1, C^2\zeta_{il}^2\right). \end{aligned} \tag{9}$$

Therefore, λ_{il}^{-1} follows an inverse Gaussian distribution

$$q(\lambda_{il}^{-1}|\boldsymbol{\eta}, \boldsymbol{\Omega}, \mathbf{R}) = \mathcal{IG}\left(\lambda_{il}^{-1}; \frac{1}{C|\zeta_{il}|}, 1\right) \tag{10}$$

where $\mathcal{IG}(x; a, b) = \sqrt{\frac{b}{2\pi x^3}} \exp\left(-\frac{b(x-1)^2}{2a^2 x}\right)$ for $a, b > 0$. We can draw a sample from an inverse Gaussian distribution in constant time (Michael et al., 1976).

4. Nonparametric Bayesian Extensions

We present an extension of the basic framework for nonparametric Bayesian inference, thus allowing to automatically resolve the model complexity in learning latent features for multi-task learning.

4.1. The Model

We again consider the simple linear discriminant function $f(\mathbf{x}; \boldsymbol{\eta}_l) = \boldsymbol{\eta}_l^\top \mathbf{x}$ and make predictions using the sign rule for multi-task binary classification. Instead of imposing a zero-mean prior on $\boldsymbol{\eta}$, which provides no structural information about what model parameters would look like, we take the suggestions by (Archambeau et al., 2011; Yang et al., 2013) and impose the structured non-zero mean matrix normal prior

$$\boldsymbol{\eta} \sim \mathcal{N}_{D,L}(\mathbf{ZV}, \mathbf{I} \otimes \mathbf{R}),$$

where both $\mathbf{Z}$ and $\mathbf{V}$ are latent and $\mathbf{Z}$ works as a projection matrix. Using such a decomposed mean would resort to a low rank approximation of correlation matrix as in (Archambeau et al., 2011; Yang et al., 2013) and provide structural information for $\boldsymbol{\eta}$. As for the task correlation matrix $\mathbf{R}$, we again assume an inverse Wishart prior with hyper-parameters $\boldsymbol{\Psi}$ and ν. We should note that the reason for choosing such a covariance for $\boldsymbol{\eta}$, as shall be seen soon, is that the modeling of the common projection matrix $\mathbf{Z}$ shared by all the tasks is essentially an indirect modeling of feature correlation matrix $\boldsymbol{\Omega}$. Thus there is no need to assume another matrix for the same use. The graphical representation is shown in Figure 1(b). In (Archambeau et al., 2011; Yang et al., 2013), some Gaussian hyper-priors are imposed on $\mathbf{Z}$ and $\mathbf{V}$, both of which are

assumed to have a finite and fixed dimension K. However, since K is an unknown parameter, a model selection procedure (e.g., cross-validation) is needed to select a good value of K. We address this issue in our Bayesian max-margin MTL by exploring the flexibility on nonparametric inference, where we let $\mathbf{Z}$ have an unbounded number of columns, corresponding to an unbounded latent feature dimensionality.

For simplicity, we start with the case where $\mathbf{Z}$ has fixed K finite columns. We impose the the matrix normal prior for $\mathbf{V}$, $\mathbf{V} \sim \mathcal{N}_{K,L}(\mathbf{0}, \mathbf{I} \otimes \mathbf{R})$. Then, we can show that the marginal prior $p_0(\boldsymbol{\eta}|\mathbf{Z}, \mathbf{R})$ by integrating out $\mathbf{V}$ is

$$p_0(\boldsymbol{\eta}|\mathbf{Z}, \mathbf{R}) = \frac{\exp\left\{-\frac{1}{2}\operatorname{Tr}\{\mathbf{R}^{-1}\boldsymbol{\eta}^\top(\mathbf{I} - \mathbf{Z}\mathbf{M}\mathbf{Z}^\top)\boldsymbol{\eta}\}\right\}}{(2\pi)^{DL/2}|\mathbf{R}|^{D/2}|\mathbf{I} + \mathbf{Z}^\top\mathbf{Z}|^{L/2}},$$

where $\mathbf{M} = (\mathbf{I} + \mathbf{Z}^\top\mathbf{Z})^{-1}$. For $\mathbf{Z}$, without loss of generality, we assume it is a binary matrix, i.e., each entry is 0 or 1. For the finite case, a simple prior is the Beta-Bernoulli prior, that is, each column k is associated with a parameter $\pi_k \sim \text{Beta}(\alpha/K, 1)$ and the entries in column k are i.i.d, namely, $z_{ik} \sim \text{Bernoulli}(\pi_k)$.

We can generalize the above process to let $\mathbf{Z}$ have an infinite number of columns. The infinite generalization of the Beta-Bernoulli prior is the hierarchical Beta process (Thibaux & Jordan, 2007) (also known as Indian buffet process, IBP (Griffiths & Ghahramani, 2005)). Although $\mathbf{Z}$ is allowed to have an infinite number of columns, it would have a finite number of non-zero, or active, columns with probability 1 under the IBP prior. We denote the matrix formed by combining these active columns as $\mathbf{Z}_+$. The column number of $\mathbf{Z}_+$ is denoted as K_+, corresponding to learned latent feature dimensionality. Then, with the infinite limit of K, the marginal prior becomes

$$p_0(\boldsymbol{\eta}|\mathbf{Z}, \mathbf{R}) = \frac{\exp\left\{-\frac{1}{2}\operatorname{Tr}\{\mathbf{R}^{-1}\boldsymbol{\eta}^\top\mathbf{Q}^{-1}\boldsymbol{\eta}\}\right\}}{(2\pi)^{DL/2}|\mathbf{R}|^{D/2}|\mathbf{I} + \mathbf{Z}_+^\top\mathbf{Z}_+|^{L/2}},$$

where we let $\mathbf{Q} = (\mathbf{I} - \mathbf{Z}_+(\mathbf{I} + \mathbf{Z}_+^\top\mathbf{Z}_+)^{-1}\mathbf{Z}_+^\top)^{-1}$ for simplicity of notations. We should note that, for all the terms related to $\boldsymbol{\eta}$, by substituting $\mathbf{Q}$ with $\boldsymbol{\Omega}$ we will get exactly the same conditional probability as in the parametric case, indicating that modeling the common projection matrix $\mathbf{Z}$ is essentially modeling the feature correlation matrix. That's why we force parts of the covariance in the prior of $\boldsymbol{\eta}$ to be identity, or else it would be redundant.

With the sampled $\boldsymbol{\eta}$, we can similarly measure the training error. We again adopt the hinge loss as a surrogate and take the expectation to account for the uncertainty of $\boldsymbol{\eta}$. This leads to the optimization problem

$$\min_{q(\boldsymbol{\eta},\mathbf{Z},\mathbf{R})\in\mathcal{P}} \mathcal{L}\big(q(\boldsymbol{\eta}, \mathbf{Z}, \mathbf{R})\big) + 2C \cdot \mathcal{R}\big(q(\boldsymbol{\eta})\big), \tag{11}$$

where $\mathcal{L}(q(\boldsymbol{\eta}, \mathbf{Z}, \mathbf{R})) = \text{KL}(q\|p_0(\boldsymbol{\eta}, \mathbf{Z}, \mathbf{R}))$.

4.2. Gibbs Sampling with Data Augmentation

By applying the same data augmentation techniques, we are able to obtain the augmented posterior distribution

$$q(\boldsymbol{\eta}, \mathbf{Z}, \mathbf{R}, \boldsymbol{\lambda}) \propto p_0(\mathbf{Z})p_0(\mathbf{R})p_0(\boldsymbol{\eta}|\mathbf{Z}, \mathbf{R})\psi(\mathbf{Y}, \boldsymbol{\lambda}|\boldsymbol{\eta}, \mathbf{X}),$$

from which we can develop a Gibbs sampler to draw samples. Specifically, the sampling probabilities for $\boldsymbol{\eta}$, $\boldsymbol{\lambda}$ and $\mathbf{R}$ are exactly the same as before except that we replace $\boldsymbol{\Omega}$ in the original probabilities by $\mathbf{Q}$.

For the binary matrix $\mathbf{Z}$, we do the sampling entry-wisely. For the existing column k, we sample z_{dk} via

$$p(z_{dk} = 1|\mathbf{Z}_{-(dk)}, \boldsymbol{\eta}, \mathbf{R}) \propto \frac{m_k - z_{dk}}{D} p(\boldsymbol{\eta}|\mathbf{Z}, \mathbf{R}), \tag{12}$$

where m_k is the number of non-zero entries in the k-th column of $\mathbf{Z}$. Note that if there exists no non-zero entries except z_{dk}, the sampler would directly remove this column. A number of new columns would be sampled via

$$p(K_{new}) \propto Poisson\left(K_{new}, \frac{\alpha}{D}\right) p(\boldsymbol{\eta}|\mathbf{Z}_{new}, \mathbf{R}), \tag{13}$$

where $\mathbf{Z}_{new}$ is obtained by appending K_{new} columns of $\mathbf{e}_d$ to $\mathbf{Z}$, where $\mathbf{e}_d$ is a 0-vector except the d-th entry being 1.

5. Multi-Task Regression

The above techniques could be generalized to tackle multi-task regression problems, where the response variable y_{il} for each instance $\mathbf{x}_{il}$ takes real values. For the i-th instance in the l-th task, we consider the linear prediction rule, $y_{il} = \boldsymbol{\eta}_l^\top \mathbf{x}_{il}$. One widely used margin-based loss measure is the ε-insensitive loss $\mathcal{R}_\varepsilon(\boldsymbol{\eta}) = \sum_{il} \max(0, |\Delta_{il}| - \varepsilon)$ for support vector regression (Smola & Schölkopf, 2004), where $\Delta_{il} = y_{il} - \boldsymbol{\eta}_l^\top \mathbf{x}_{il}$ is the margin. To do Bayesian max-margin learning, we define the expected ε-intensive loss as $\mathcal{R}_\varepsilon(q(\boldsymbol{\eta})) = \sum_{il} \mathbb{E}_q(\max(0, |\Delta_{il}| - \varepsilon))$. Then we define BM-MTL regression model as solving the entropy-regularized loss minimization problem

$$\min_{q(\boldsymbol{\eta})\in\mathcal{P}} \mathcal{L}\big(q(\boldsymbol{\eta})\big) + 2C \cdot \mathcal{R}_\varepsilon\big(q(\boldsymbol{\eta})\big).$$

The resulting posterior probability follows the form as in Eq. (6), with the pseudo-likelihood being $\psi(\mathbf{Y}|\boldsymbol{\eta}, \mathbf{X}) \propto \prod_{l=1}^{L}\prod_{i=1}^{N_l} \exp\{-2C\max(0, |\Delta_{il}| - \varepsilon)\}$.

By noting that $\max(0, |x| - \varepsilon) = \max(0, x - \varepsilon) + \max(0, -x - \varepsilon)$, we can do similar data augmentation and express the unnormalized likelihood for each instance $(\mathbf{x}_{il}, y_{il})$ as

$$\psi(y_{il}|\boldsymbol{\eta}, \mathbf{x}_{il}) = \int_0^\infty \int_0^\infty \psi(y_{il}, \lambda_{il}, \omega_{il}|\boldsymbol{\eta}, \mathbf{x}_{il}) \mathrm{d}\lambda_{il}\mathrm{d}\omega_{il},$$

where $\psi(y_{il}, \lambda_{il}, \omega_{il}|\boldsymbol{\eta}, \mathbf{x}_{il}) = \frac{1}{\sqrt{2\pi\lambda_{il}}} \exp\{-\frac{1}{2\lambda_{il}}(\lambda_{il} + C(\Delta_{il} - \varepsilon))^2\} \times \frac{1}{\sqrt{2\pi\omega_{il}}} \exp\{-\frac{1}{2\omega_{il}}(\omega_{il} - C(\Delta_{il} + \varepsilon))^2\}$,

with the augmented variables $\boldsymbol{\lambda} = \{\{\lambda_{il}\}_{i=1}^{N_l}\}_{l=1}^{L}$ and $\boldsymbol{\omega} = \{\{\omega_{il}\}_{i=1}^{N_l}\}_{l=1}^{L}$. Then, the target posterior distribution $q(\boldsymbol{\eta}, \boldsymbol{\Omega}, \mathbf{R})$ is the marginal of the complete posterior

$$q(\boldsymbol{\eta}, \boldsymbol{\Omega}, \mathbf{R}, \boldsymbol{\lambda}, \boldsymbol{\omega}) \propto p_0(\boldsymbol{\Omega}) p_0(\mathbf{R}) p_0(\boldsymbol{\eta}|\boldsymbol{\Omega}, \mathbf{R}) \psi(\mathbf{Y}, \boldsymbol{\lambda}, \boldsymbol{\omega}|\boldsymbol{\eta}, \mathbf{X}),$$

where $\psi(\mathbf{y}, \boldsymbol{\lambda}, \boldsymbol{\omega}|\boldsymbol{\eta}, \mathbf{X}) = \prod_l \prod_i \psi(y_{il}, \lambda_{il}, \omega_{il}|\boldsymbol{\eta}, \mathbf{x}_{il})$. Following similar derivations as in the classification models, the Gibbs sampler iteratively performs the steps:

For $\boldsymbol{\eta}$: again we can employ an effective sampling of each $\boldsymbol{\eta}_l$ task-wisely. With the matrix normal prior on $\boldsymbol{\eta}$, the conditional probability is still Gaussian:

$$q(\boldsymbol{\eta}_l|\boldsymbol{\Omega}, \mathbf{R}, \boldsymbol{\lambda}, \boldsymbol{\eta}_{-l}) \propto \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l), \tag{14}$$

where the posterior covariance matrix and mean are computed as $\boldsymbol{\Sigma}_l = (\sum_i C^2 \rho_{il} \mathbf{x}_i \mathbf{x}_i^\top + R_{ll}^{-1} \Omega^{-1})^{-1}$ and $\boldsymbol{\mu}_l = \boldsymbol{\Sigma}_l(-\frac{1}{2}\sum_{k \neq l} \mathbf{R}_{lk}^{-1} \boldsymbol{\Omega}^{-1} \boldsymbol{\eta}_k + \sum_i C^2(y_{il}\rho_{il} - \varepsilon\sigma_{il})\mathbf{x}_i)$ with $\rho_{il} = (\frac{1}{\lambda_{il}} + \frac{1}{\omega_{il}})$ and $\sigma_{il} = (\frac{1}{\lambda_{il}} - \frac{1}{\omega_{il}})$.

For $\boldsymbol{\lambda}$ and $\boldsymbol{\omega}$: the conditional distribution for both sets of augmented variables are still inverse Gaussian:

$$\begin{aligned} q(\lambda_{il}^{-1}|\boldsymbol{\eta}, \boldsymbol{\Omega}, \mathbf{R}) &= \mathcal{IG}\left(\lambda_{il}^{-1}; \frac{1}{C|\Delta_{il} - \varepsilon|}, 1\right), \\ q(\omega_{il}^{-1}|\boldsymbol{\eta}, \boldsymbol{\Omega}, \mathbf{R}) &= \mathcal{IG}\left(\omega_{il}^{-1}; \frac{1}{C|\Delta_{il} + \varepsilon|}, 1\right). \end{aligned} \tag{15}$$

The sampling probabilities for $\boldsymbol{\Omega}$ and $\mathbf{R}$ are exactly the same as Eq. (8) in the classification case. We could futher derive the sampling probabilities for NPBM-MTL model by substituting $\boldsymbol{\Omega}$ with $\mathbf{Q}$ in sampling distributions of the parametric regression case, together with the sampling probability of $\mathbf{Z}$ following Eq. (12) and Eq. (13).

6. Experiments

We present empirical studies for both multi-task classification and multi-task regression.

6.1. Multi-Task Classification

6.1.1. Datasets, Methods and Setups

We use four multi-label datasets that are publicly available[1]. Table 1 summarizes their statistics. For multi-label classification, we formulate it as a multi-task learning problem, where each task is a binary classifier determining whether an instance has a particular label.

We compare with various baselines, including LR (Independent Logistic Regression), MTL-C (Clustered Multi-task Learning) (Jacob et al., 2008), MTL-F (Multi-task Feature Learning) (Argyriou et al., 2006) and MT-iLSVM (Multi-task infinite latent SVMs) (Zhu et al., 2011). MTL-C could learn clusters of different tasks to capture the task correlations. MTL-F learns a low-dimensional representation shared across a set of related tasks, while MT-iLSVM learns an infinite dimensional feature representation. Both MTL-F and MT-iLSVM are essentially learning the feature correlations. In BM-MTL, we have the flexibility to learn the correlation matrices or simply set them to some fixed matrices. There are also various choices for hyper-parameters $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$ in inverse Wishart priors. We either set $\mathbf{R}$ or $\boldsymbol{\Omega}$ to be fixed identity or sample them with $\boldsymbol{\Psi}_1$ or $\boldsymbol{\Psi}_2$ being identity, yielding four different configurations of BM-MTL listed below:

- (**I**-$\boldsymbol{\Omega}$&**I**-$\mathbf{R}$) We fix both $\boldsymbol{\Omega}$ and $\mathbf{R}$ to be identity.
- ($\boldsymbol{\Omega}$&**I**-$\mathbf{R}$) We fix $\mathbf{R}$ to be identity and learn $\boldsymbol{\Omega}$ with $\boldsymbol{\Psi}_1$ being identity.
- (**I**-$\boldsymbol{\Omega}$&$\mathbf{R}$) We fix $\boldsymbol{\Omega}$ to be identity and learn $\mathbf{R}$ with $\boldsymbol{\Psi}_2$ being identity.
- ($\boldsymbol{\Omega}$&$\mathbf{R}$) We learn both $\boldsymbol{\Omega}$ and $\mathbf{R}$ with $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$ being identities.

Finally, for the nonparametric NPBM-MTL, we learn both $\mathbf{Z}$ and $\mathbf{R}$ with $\boldsymbol{\Psi}$ being identity. The regularization parameter C is chosen from $[10^{-3}, 10^3]$ using 5-fold cross validation. We use the F1-score as the performance measure, which accounts for label bias in the datasets. To avoid the condition that the sampled $\boldsymbol{\Omega}^{-1}$ and $\mathbf{R}^{-1}$ are too large or close to zero, we divide the sampled matrices by a constant number and force the diagonal entries to be at least one.

Table 1. Statistics of various datasets, where "Avg-labels" stands for the average number of labels per instance.

Datasets	Num Samples	L	D	Avg-labels
Emotions	593	6	72	1.869
Cal500	502	174	68	26.044
Scene	2,407	6	294	1.074
Yeast	2,417	14	103	4.237

6.1.2. Results and Analysis

Table 2 shows the mean F1-scores and standard deviation (in parentheses) obtained from runs on 5 random splits of each dataset, where each row corresponds to a model and each column corresponds to a dataset. Our random splits are obtained following the datasets' original training-testing split ratios, which are provided along with the data.

We observe that among the four BM-MTL models with various configurations, the one with ($\boldsymbol{\Omega}$&$\mathbf{R}$) provides consistently promising results — it either offers the best performance or gives a competitive one. This shows the benefits brought by simultaneously estimating task correlations and feature correlations. As for the two non-parametric methods, NPBM-MTL obtains significant improvements over MT-iLSVM by modeling task correlations and relieving

[1] http://mulan.sourceforge.net/datasets.html

Table 2. Results for different models on the four datasets. Competitive results are shown in bold, and the best one is marked with "*".

MODELS	F1(%)	F1(%)	F1(%)	F1(%)
	EMOTIONS	CAL500	SCENE	YEAST
LR	54.83 (1.63)	32.05 (1.63)	61.65 (1.45)	61.44 (0.92)
MTL-C	63.50 (1.67)	33.51 (1.39)	64.98 (0.66)	61.86 (0.69)
MTL-F	64.51 (1.10)	34.01 (1.47)	64.23 (0.89)	62.66 (0.55)
BM-MTL(**I**-$\mathbf{\Omega}$&**I**-**R**)	64.16 (1.19)	34.65 (1.06)	65.85 (0.78)	63.75 (0.57)
BM-MTL($\mathbf{\Omega}$&**I**-**R**)	**65.22 (1.34)**	34.91 (1.05)	65.97 (0.70)	**64.20 (0.59)**
BM-MTL(**I**-$\mathbf{\Omega}$&**R**)	**64.97 (1.37)**	**35.07 (1.22)**	**66.30 (0.70)**	**64.16 (0.53)**
BM-MTL($\mathbf{\Omega}$&**R**)	**65.67 (1.32)***	**35.54 (1.09)**	**66.33 (0.70)**	**64.36 (0.58)***
MT-iLSVM	62.27(2.25)	32.20(1.09)	62.27(2.46)	61.81(0.60)
NPBM-MTL(**Z**&**R**)	**65.53 (1.34)**	**35.67 (0.69)***	**66.85 (0.55)***	**64.21 (0.44)**

the truncated mean-field assumption made by MT-iLSVM, which only estimates the feature correlations. Overall, both BM-MTL and NPBM-MTL could significantly improve the performance over existing competitors[2], mainly due to the advantages brought by conjoining Bayesian methods and max-margin learning as well as the joint estimation of task correlations and feature correlations.

We also compare with the methods in (Archambeau et al., 2011). Since their code is not available, we compare with the reported accuracy on the *yeast* and *scene* datasets. Our NPBM-MTL achieves an accuracy of 80.01 (0.06) on *yeast*, slightly better than the reported accuracy of 79.88 (0.17) in (Archambeau et al., 2011), and gives competitive accuracy of 88.95 (0.07) on *scene* where the reported result is 88.97 (0.34). Our model could not only give stabler performance, but also free users from manually tuning the latent feature dimensionality, contrast to methods in (Archambeau et al., 2011).

6.2. Multi-Task Regression

6.2.1. DATASETS, METHODS AND SETUPS

We use the public *School* dataset, which consists of the examination records of 15,362 students from 139 secondary schools in years 1985, 1986 and 1987. The dataset has been used to study the effectiveness of schools. It has been used extensively to evaluate multi-task learning methods (Bakker & Heskes, 2003; Zhang & Yeung, 2012), where the goal of the task is to predict the exam scores of students from different schools based on four student-dependent features and four school-dependent features. To make a fair comparison, we follow the same setup described in (Bakker & Heskes, 2003) and use the same 10 random splits of the data to do the training and testing. The average number of students included in training set is about 80 per school, while the number in testing set is about 30.

We compare with the state-of-the-art results of BMTL (Bayesian Multi-task Learning) (Bakker & Heskes, 2003), MTGP (Multi-task Gaussian Processes) (Bonilla et al., 2007), MTRL (Convex Multi-task Relationship Learning) (Zhang & Yeung, 2012), STL (Single Task Learning) as reported in (Bonilla et al., 2007; Zhang & Yeung, 2012) and the MT-iLSVM regression model (Zhu et al., 2011). We empirically set ε to be 0.001. The hyper-parameter of the IBP prior of $\mathbf{Z}$ (i.e., α) is fixed at 5 in this experiment, and we will return to investigate the sensitivity over this parameter in Section 6.3.2. We search the regularization parameter C in the range of [0.1,10] with 5-fold cross validation.

Table 3. Experimental results for different models on School dataset. PEV stands for the percentage of explained variance.

BASELINES	PEV(%)	MODEL	PEV(%)
STL	23.5 (1.9)	BM-MTL(**I**-$\mathbf{\Omega}$&**I**-**R**)	31.12 (1.02)
BMTL	29.5 (0.4)	BM-MTL($\mathbf{\Omega}$&**I**-**R**)	31.21 (1.02)
MTGP	29.2 (1.6)	BM-MTL(**I**-$\mathbf{\Omega}$&**R**)	31.54 (1.04)
MTRL	29.9 (1.8)	BM-MTL($\mathbf{\Omega}$&**R**)	31.56 (1.09)
MT-iLSVM	31.13 (1.15)	NPBM-MTL(**Z**&**R**)	**31.86 (1.00)**

For performance measure, we use *percentage of explained variance* (PEV) (Bakker & Heskes, 2003). PEV is defined as the total variance of the data minus the sum-squared error on the test set as a percentage of the total variance.

6.2.2. RESULTS AND ANALYSIS

Table 3 shows the results. We can see that both BM-MTL and NPBM-MTL could work very well on regression tasks. For the four configurations of BM-MTL, the hybrid learning of $\mathbf{\Omega}$ and $\mathbf{R}$ again yields the best performance. Although all these four configurations could outperform other state-of-the-art methods, the variances are large compared to some of the baselines. While for NPBM-MTL, it not only gives the best performance, but also enjoys a small variance. This may result from the modeling of latent features — when the original features are redundant and noisy, the learned latent feature dimensionality would be low, which could work as a refiner of data noise and result in simple and effective models. In Section 6.3.2, we will see that the learned latent feature dimensionality is low on *School*, which gives evidence to our statement.

[2]We performed the 2-tailed t-test. For BM-MLT, the p-values over all the existing methods on all the datasets are less than 0.027; similarly for NPBM-MTL.

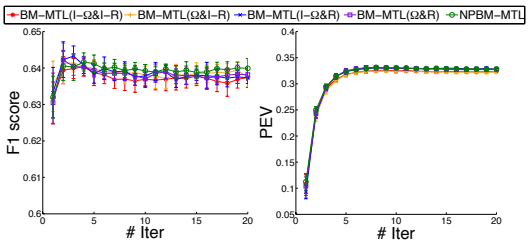

Figure 2. Prediction performance w.r.t the number of iterations (i.e., burn-in steps) on (a) *yeast* and (b) *School* datasets.

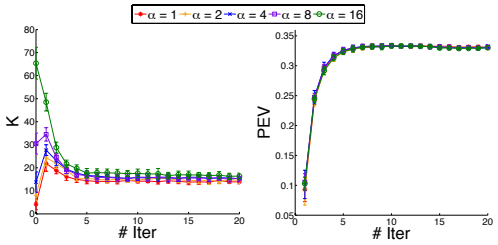

Figure 3. Impact of α on (a) learned latent feature dimensions and (b) performance of NPBM-MTL.

6.3. More Discussions

6.3.1. Burn-in Steps

We analyze the effects of the number of burn-in steps on the performance for both multi-task classification and regression problems. For classification, we choose the *yeast* dataset as an example, and for regression we choose one of the 10 random splits of the *school* data. Figure 2 presents the performance changes of both the parametric BM-MTL and nonparametric NPBM-MTL with respect to the number of burn-in steps. We can observe that all our methods converge quickly. For example, on both datasets, we need about 10 burn-in steps to get stable prediction performance.

6.3.2. Sensitivity Analysis and Latent Features

Since the initialization of $\mathbf{Z}$ follows an IBP, the number of new columns added to $\mathbf{Z}$ follows Poisson($\frac{\alpha}{i}$) when initializing the i-th row of $\mathbf{Z}$, where α is the IBP hyper-parameter. Thus the initial latent dimensions of $\mathbf{Z}$ would be different if using different α values, resulting in initial latent dimensions that may be far from the actual one. To analyze whether NPBM-MTL is sensitive to such an initialization, we analyze the influence of α on both the learned feature dimensionality and prediction performance on the *School* dataset. We set α to be $2^{[0:4]}$. Figure 3 shows the results.

We can see that even though the initial dimensions vary a lot with different α, with no more than 10 iterations, all the NPBM-MTL models with different α values converge to a similar dimension, which is about 15. Meanwhile, the prediction performance gradually improves until the sampling procedure converges and the latent feature dimensionality settles down. This shows that NPBM-MTL is able to converge quickly and is stable against different choices of α.

Table 4. Learned latent feature dimensionality K_+ on different datasets. D is the original feature dimensionality.

	Emotions	Cal500	Scene	Yeast	School
K_+	57.8 (2.0)	4.4 (0.5)	48.2 (1.9)	4.8 (0.8)	14.8 (1.1)
D	72	68	294	103	28

We also observe in Table 4 that the inferred latent feature dimensions are generally much smaller than the original feature dimensions. For example, the original feature dimensionality in the *School* dataset is nearly 30, but the learned latent feature dimensionality is only about 15. On most datasets, the learned K_+ is relatively small compared to D, thus enjoying a denoising effect and resulting in stabler performance (except on Emotions, whose K_+ is close to D, and the performance on Emotions, as can be seen from Table 2, is not that stable). The low ratio of K_+/D indicates that usually data has redundant features, and by applying our non-parametric method we would be able to learn the "true" feature dimensionality automatically instead of setting them manually in the model.

7. Conclusions and Future Work

We present Bayesian max-margin multi-task learning, which conjoins the max-margin learning with Bayesian formalism and allows discriminative max-margin models to enjoy the great flexibility of Bayesian methods on incorporating rich prior information and performing nonparametric Bayesian inference to learn latent features or structures as well as the latent feature dimensionality. We present simple Gibbs sampling algorithms by exploring data augmentation techniques. Our algorithms are truncation-free and assumption-free when applied to nonparametric Bayesian models. Experimental results demonstrate promising prediction performance over various competitors.

Bayesian max-margin models for multi-task learning could be further extended. For example, both BM-MTL and NPBM-MTL may be expensive in training when the feature dimension is huge, because the sampling of weight matrices would involve large matrix inversion. To tackle such a problem, feature-dimension reduction techniques may be used. Another interesting direction is to selectively share information among tasks (Kumar & III, 2012), which may be beneficial especially when we have many tasks. In theory, we can consider the selective sharing structure by formulating it via some prior. We leave these for future work.

Acknowledgments

The work is supported by the National Basic Research Program of China (No. 2013CB329403), National Natural Science Foundation of China (Nos. 61322308, 61332007), and Tsinghua University Initiative Scientific Research Program (No. 20121088071).

References

Archambeau, C., Guo, S., and Zoeter, O. Sparse Bayesian multi-task learning. In *NIPS*, pp. 1755–1763, 2011.

Argyriou, A., Evgeniou, T., and Pontil, M. Multi-task feature learning. In *NIPS*, pp. 41–48, 2006.

Bakker, B. and Heskes, T. Task clustering and gating for Bayesian multitask learning. *JMLR*, 4:83–99, 2003.

Barndorff-Nielsen, O., Blæsild, P., Jensen, J., and Jorgensen, B. Exponential transformation models. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 379:41–65, 1982.

Bonilla, E., Chai, K., and Williams, C. Multi-task Gaussian process prediction. In *NIPS*, 2007.

Caruana, R. Multitask learning. *Machine Learning*, 28: 41–75, 1997.

Chen, J., Tang, L., and Liu, J.and Ye, J. A convex formulation for learning shared structures from multiple tasks. In *ICML*, pp. 137–144, 2009.

Devroye, L. *Non-uniform random variate generation*. Springer-Verlag, 1986.

Germain, P., Lacasse, A., Laviolette, F., and Marchand, M. PAC-Bayesian learning of linear classifiers. In *ICML*, pp. 353–360, 2009.

Griffiths, T. and Ghahramani, Z. Infinite latent feature models and the Indian Buffet Process. In *NIPS*, 2005.

Hariharan, B., Zelnik-Manor, L., Vishwanathan, S. V. N., and Varma, M. Large scale max-margin multi-label classification with priors. In *ICML*, pp. 423–430, 2010.

Jaakkola, T., Meila, M., and Jebara, T. Maximum entropy discrimination. In *NIPS*, 1999.

Jacob, L., Bach, F., and Vert, J. Clustered multi-task learning: A convex formulation. In *NIPS*, pp. 745–752, 2008.

Koyejo, O. and Ghosh, J. Constrained bayesian inference for low rank multitask learning. *UAI*, pp. 97–106, 2013.

Kumar, A. and III, Daume H. Learning task grouping and overlap in multi-task learning. *ICML*, pp. 690–697, 2012.

Mardia, K., Kent, J., and Bibby, J. *Multivariate analysis*. Academic press, 1980.

McAllester, D. PAC-Bayesian stochastic model selection. *Machine Learning*, 51:5–21, 2003.

Michael, J.R., Schucany, W.R., and Haas, R.W. Generating random variates using transformations with multiple roots. *The American Statistician*, 30(2):88–90, 1976.

Obozinski, G., Taskar, B., and Jordan, M. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20: 231–252, 2010.

Polson, N. and Scott, S. Data augmentation for support vector machines. *Bayesian Analysis*, 6:1–24, 2011.

Rai, P. and Daume, H. Infinite predictor subspace models for multitask learning. In *AISTATS*, pp. 613–620, 2010.

Smola, A. and Schölkopf, B. A tutorial on support vector regression. *Statistics and Computing*, 14:199–222, 2004.

Tanner, M. and Wong, W. The calculation of posterior distributions by data augmentation. *Journal of the Americal Statistical Association (JASA)*, 82(398):528–540, 1987.

Thibaux, R. and Jordan, M. Hierarchical Beta processes and the Indian Buffet Process. In *AISTATS*, pp. 564–571, 2007.

Thrun, S. and O'Sullivan, J. Discovering structure in multiple learning tasks: The TC algorithm. In *ICML*, pp. 489–497, 1996.

Widmer, C. and Rätsch, G. Multitask learning in computational biology. *JMLR*, 27:207–216, 2012.

Xue, Y., Liao, X., Carin, L., and Krishnapuram, B. Multi-task learning for classification with Dirichlet process priors. *JMLR*, 8:35–63, 2007.

Yang, M., Li, Y., and Zhang, Z. Multi-task learning with Gaussian matrix generalized inverse Gaussian model. In *ICML*, pp. 423–431, 2013.

Yu, S., Tresp, V., and Yu, K. Robust multi-task learning with *t*-processes. In *ICML*, pp. 1103–1110, 2007.

Zhang, T., Ghanem, B., Liu, S., and Ahuja, N. Robust visual tracking via multi-task sparse learning. In *CVPR*, pp. 2042–2049, 2012.

Zhang, Y. and Schneider, J. Learning multiple tasks with a sparse matrix-normal penalty. In *NIPS*, pp. 2550–2558, 2010.

Zhang, Y. and Yeung, D. A convex formulation for learning task relationships in multi-task learning. *arXiv:1203.3536*, 2012.

Zhu, J., Chen, N., and Xing, E. Infinite latent SVM for classification and multi-task learning. In *NIPS*, pp. 1620–1628, 2011.

Zhu, J., Zheng, X., Zhou, L., and Zhang, B. Scalable inference in max-margin topic models. In *KDD*, pp. 964–972, 2013.

Zhu, J., Chen, N., Perkins, H., and Zhang, B. Gibbs max-margin topic models with data augmentation. *JMLR*, 15:1073–1110, 2014a.

Zhu, J., Chen, N., and Xing, E. Bayesian inference with posterior regularization and applications to infinite latent SVMs. *JMLR (in press, arXiv:1210.1766v3)*, 2014b.