A Probabilistic Model for Dirty Multi-task Feature Selection

Daniel Hernández-Lobato DANIEL.HERNANDEZ@UAM.ES
Universidad Autónoma de Madrid, Computer Science Department, Madrid, 28049, Spain

José Miguel Hernández-Lobato JMH@SEAS.HARVARD.EDU
Harvard University, School of Engineering and Applied Sciences, Cambridge, MA 02138, USA

Zoubin Ghahramani ZOUBIN@ENG.CAM.AC.UK
University of Cambridge, Department of Engineering, Cambridge CB2 1PZ, UK

Abstract

Multi-task feature selection methods often make the hypothesis that learning tasks share relevant and irrelevant features. However, this hypothesis may be too restrictive in practice. For example, there may be a few tasks with specific relevant and irrelevant features (outlier tasks). Similarly, a few of the features may be relevant for only some of the tasks (outlier features). To account for this, we propose a model for multi-task feature selection based on a robust prior distribution that introduces a set of binary latent variables to identify outlier tasks and outlier features. Expectation propagation can be used for efficient approximate inference under the proposed prior. Several experiments show that a model based on the new robust prior provides better predictive performance than other benchmark methods.

1. Introduction

When the number of samples is smaller or equal to the number of attributes or features, regression problems are under-determined. In this case, a linear model is too complex to explain the observed data since an infinite number of model coefficients perfectly fit the data. In this context, sparsity, *i.e.*, the assumption of zeros in the model coefficients, plays a strong regularization role that can be useful to obtain estimates with good generalization properties. Sparsity can be favored by using sparsity enforcing priors in probabilistic models or by optimizing a loss function penalized by a sparsity inducing norm (Carvalho et al., 2009; Jalali et al., 2010; Vogt & Roth, 2010). The assumption of zeros in the model coefficients is equivalent to the assumption of only a few of relevant features for prediction.

Multi-task feature selection methods are used to improve the process of inferring the model coefficients from the observed data under the sparsity assumption (Vogt & Roth, 2010; Hernández-Lobato et al., 2010; Obozinski et al., 2009; Xiong et al., 2007; Zhang et al., 2008). In these methods several tasks that have a common feature space are solved simultaneously, often under the assumption that the tasks share relevant and irrelevant features, as illustrated by Figure 2 (top). However, in some situations this hypothesis may be too restrictive (Jalali et al., 2010). As illustrated by Figure 2 (bottom), a few of the tasks may have specific relevant / irrelevant features (outlier tasks), and a few of the features may be arbitrarily relevant / irrelevant across tasks (outlier features). In this situation, traditional multi-task feature selection methods are expected to perform poorly.

In this paper we propose a multi-task feature selection model, based on a robust prior distribution, that is expected to have better properties in the presence of diverse tasks, *i.e.*, data with the properties described above. Exact inference is intractable in such a model. However, expectation propagation can be used for efficient approximate inference (Minka, 2001). Several experiments involving the reconstruction of gene regulatory networks, the denoising of natural images and the prediction of drug sensitivity from microarray data illustrate the benefits of the model proposed. Specifically, it has better prediction properties than other methods from the literature and it can be used to successfully identify relevant attributes for prediction, alongside with outlier tasks and features, which may be useful to better understand the characteristics of the observed data.

2. Dirty Multi-task Feature Selection

Assume K regression tasks with data $\{\mathbf{X}^{(k)}, \mathbf{y}^{(k)}\}_{k=1}^{K}$, where $\mathbf{X}^{(k)}$ and $\mathbf{y}^{(k)}$ are the design matrix and the vector

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

of targets for task k, respectively. All tasks share the same d attributes or features, but feature values can be different across tasks. A linear model is considered for each task, *i.e.*, $\mathbf{y}^{(k)} = \mathbf{X}^{(k)}\mathbf{w}^{(k)} + \boldsymbol{\epsilon}^{(k)}$, where $\mathbf{w}^{(k)} \in \mathbb{R}^d$ is the vector of model coefficients for task k and $\boldsymbol{\epsilon}^{(k)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma^2_{(k)})$ is Gaussian noise with variance $\sigma^2_{(k)}$. Let $\mathbf{W}$ be a $K \times d$ matrix whose k-th row is $\mathbf{w}^{(k)}$ and $\mathbf{Y}$ a matrix whose k-th row is $\mathbf{y}^{(k)}$. Define $\boldsymbol{\mathcal{X}} = \{\mathbf{X}^{(k)}\}_{k=1}^K$ and $\boldsymbol{\sigma}^2 = \{\sigma^2_{(k)}\}_{k=1}^K$. The likelihood for $\mathbf{W}$ and $\boldsymbol{\sigma}^2$ is:

$$p(\mathbf{Y}|\boldsymbol{\mathcal{X}}, \mathbf{W}, \boldsymbol{\sigma}^2) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{y}^{(k)}|\mathbf{X}^{(k)}\mathbf{w}^{(k)}, \mathbf{I}\sigma^2_{(k)}) \,. \quad (1)$$

Moreover, feature selection for each task, or equivalently, sparsity in each $\mathbf{w}^{(k)}$ is expected to be beneficial. We also assume that the K tasks share relevant and irrelevant features, but we allow for small *deviations* from this hypothesis. All this prior knowledge is introduced in the model by a robust prior for $\mathbf{W}$ described in the next section.

2.1. Robust prior distribution

The prior considered is based on the discrete mixture prior introduced in (Carvalho et al., 2009). Thus, we first describe and motivate the use of that prior to favor sparse solutions. Then, we show how it can be extended to perform feature selection across several tasks in a robust way.

2.1.1. Discrete mixture prior

This is a *spike and slab* prior in which the i-th coefficient of task k satisfies $w_i^{(k)} \sim (1-\rho)\delta_0 + \rho\pi(w_i^{(k)})$, where ρ is the prior inclusion probability, δ_0 is a point of probability mass at zero, and $\pi(\cdot)$ is a density that specifies the distribution of the coefficients that are not zero. Each $w_i^{(k)}$ is *a priori* zero with probability $(1-\rho)$. In (Carvalho et al., 2009) it is suggested for $\pi(\cdot)$ the Strawderman-Berger distribution (Strawderman, 1971; Berger, 1980), which has Cauchy-like tails and yet allows for a closed form convolution with a Gaussian likelihood. This distribution is a scale mixture of Gaussians (Armagan et al., 2011) defined as:

$$\begin{aligned} \pi\left(w_i^{(k)}\right) &= \int \mathcal{N}(w_i^{(k)}|0, \lambda_i^2) \frac{\lambda_i}{(\lambda_i^2+1)^{\frac{3}{2}}} \, d\lambda_i \\ &= \frac{1}{\sqrt{2\pi}} \left(1 - |w_i^{(k)}| \frac{\Phi(-|w_i^{(k)}|)}{\mathcal{N}(w_i^{(k)}|0,1)}\right) , \end{aligned} \quad (2)$$

where $|\cdot|$ denotes absolute value, $\Phi(\cdot)$ is the cdf of a standard Gaussian distribution, $\mathcal{N}(\cdot|0,1)$ is the standard Gaussian density, and $\lambda_i/(\lambda_i^2+1)^{\frac{3}{2}}$ is the density assumed for λ_i. Figure 1 (left and middle) compares the discrete mixture prior with other priors from the literature (an arrow means a point of probability mass). We observe that the discrete mixture has heavy tails to explain coefficients that significantly differ from zero. It also has a point mass at zero that allows for exact zeros in the coefficients.

Let $\sigma^2_{(k)} = 1$ and $\mathbf{X}^{(k)} = \mathbf{I}$, and define $\kappa_i = 1/(1+\lambda_i^2)$. Carvalho et al. (2009) shows that the posterior mean for $w_i^{(k)}$ is in this case $(1-\kappa_i)y_i^{(k)}$, where κ_i is a random shrinkage coefficient. Figure 1 (right) displays the prior density for κ_i that results from each prior for $w_i^{(k)}$. The prior for κ_i is obtained by applying the change of variables $\kappa_i = 1/(1+\lambda_i^2)$ to the prior for λ_i, which in the case of the discrete-mixture prior is a mixture between the distribution for λ_i assumed in (2) and a point mass at zero. Figure 1 (right) shows that under the discrete-mixture prior *a priori* we expect to observe $\kappa_i = 1$ as a consequence of the point mass at one in the prior for κ_i. Furthermore, we also expect to observe $\kappa_i \approx 0$ as a consequence of the density tending to infinity at zero. These two values for κ_i correspond respectively to total shrinkage (zero values) and no shrinkage at all (non-zero values) for $w_i^{(k)}$. By contrast, under the other priors for $w_i^{(k)}$ the density for κ_i tends to zero at zero (Laplace) or tends to zero at one (Student's T). This means that these priors will shrink relevant coefficients and will not shrink irrelevant coefficients, respectively. Thus, the discrete mixture prior can be considered as a *golden standard* for learning under sparsity (Carvalho et al., 2009).

2.1.2. Extension of the discrete-mixture prior

The previous prior is extended to perform feature selection across several tasks. We assume that the tasks share in general relevant and irrelevant features, but we consider a few outlier tasks with specific relevant / irrelevant features and a few outlier features that may be arbitrarily relevant / irrelevant for each task. This is illustrated in Figure 2 (bottom). Tasks 4 and 8 are outlier tasks and features 19 and 21 are outlier features. All other tasks and features share the hypothesis of jointly relevant / irrelevant features across tasks.

To model this type of prior knowledge we introduce the following binary latent variables:

- z_i Indicates whether feature i is an outlier ($z_i = 1$) or not ($z_i = 0$). If it is an outlier it can be independently relevant or irrelevant for each task.
- ω_k Indicates whether task k is an outlier ($\omega_k = 1$) or not ($\omega_k = 0$). If it is an outlier it can have specific relevant and irrelevant features for prediction.
- γ_i Indicates whether the non-outlier feature i is relevant ($\gamma_i = 1$) for prediction or not ($\gamma_i = 0$) in all tasks that are not outliers, *i.e.*, those tasks for which $\omega_k = 0$.
- $\tau_i^{(k)}$ Indicates whether, given that task k is an outlier task, *i.e.*, $\omega_k = 1$, feature i for that task is relevant ($\tau_i^{(k)} = 1$) or irrelevant ($\tau_i^{(k)} = 0$) for prediction.
- $\eta_i^{(k)}$ Indicates whether, given that feature i is an outlier feature,

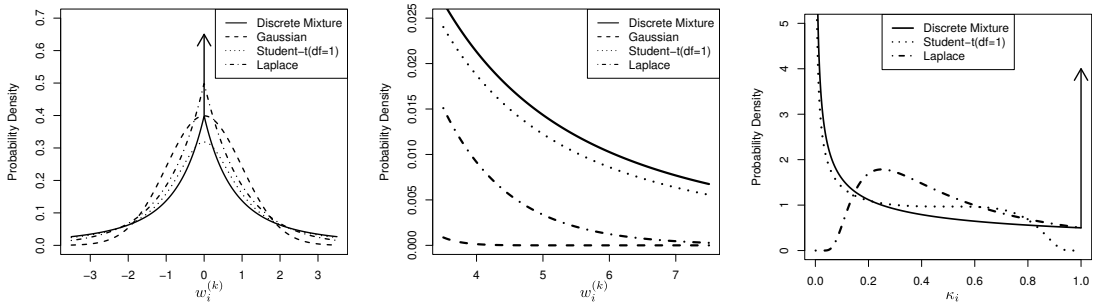

Figure 1. (left) Density of different priors. Note the spike of the discrete mixture at the origin. (middle) Tails of the different priors. (right) Prior density of the shrinkage parameter κ_i for the discrete mixture prior and for other priors from the literature.

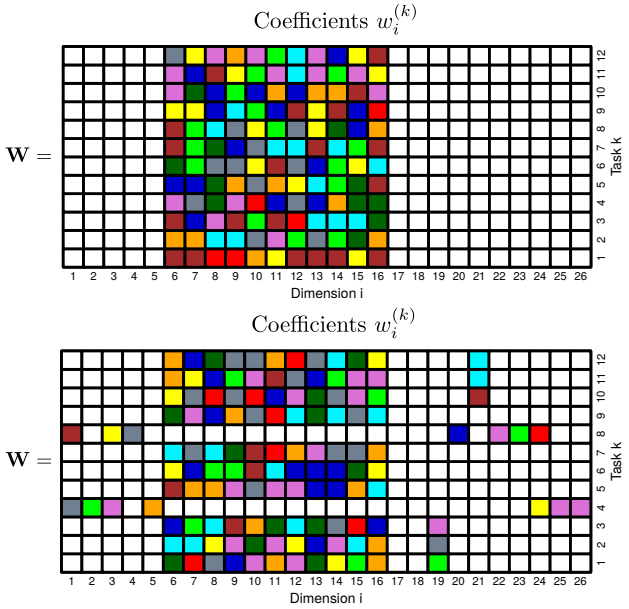

Figure 2. (top) Traditional multi-task feature selection: All tasks share relevant and irrelevant features (model coefficients). (bottom) Dirty multi-task feature selection: Most tasks share relevant and irrelevant features, but we allow for outlier tasks (tasks 4 and 8) and for outlier features (dimensions 19 and 21). White squares represent irrelevant coefficients that are equal to zero. Colored squares represent relevant coefficients with non-zero values.

that particular feature is relevant for prediction in task k ($\eta_i^{(k)} = 1$) or not ($\eta_i^{(k)} = 0$).

Let $\mathbf{\Omega}$ be the collection of all these latent variables, *i.e.* $\mathbf{\Omega} = \{\mathbf{z}, \boldsymbol{\omega}, \boldsymbol{\gamma}, \{\boldsymbol{\tau}^{(k)}\}_{k=1}^K, \{\boldsymbol{\eta}^{(k)}\}_{k=1}^K\}$. Given the latent variables we can specify the prior distribution for $\mathbf{W}$:

$$p(\mathbf{W}|\mathbf{\Omega}) = \prod_{i=1}^{d}\prod_{k=1}^{K} p(w_i^{(k)}|\mathbf{\Omega})\,, \tag{3}$$

where $p(w_i^{(k)}|\mathbf{\Omega}) = \{\pi(w_i^{(k)})^{\eta_i^{(k)}}\delta_0^{1-\eta_i^{(k)}}\}^{z_i}\{[\pi(w_i^{(k)})^{\tau_i^{(k)}}\delta_0^{1-\tau_i^{(k)}}]^{\omega_k}[\pi(w_i^{(k)})^{\gamma_i}\delta_0^{1-\gamma_i}]^{1-\omega_k}\}^{1-z_i}$. Under this prior a coefficient $w_i^{(k)}$ is different from zero if (i) it corresponds to an outlier feature ($z_i = 1$) relevant for task k ($\eta_i^{(k)} = 1$); or (ii) it does not correspond to an outlier feature ($z_i = 0$), but it corresponds to an outlier task ($\omega_k = 1$) and the feature is relevant for that task ($\tau_i^{(k)} = 1$); or (iii) it does not correspond to an outlier feature ($z_i = 0$), nor an outlier task ($\omega_k = 0$), but the feature is relevant for prediction across tasks ($\gamma_i = 1$). Otherwise, the coefficient is zero.

The hyper-priors for the latent variables are Bernoullis with parameters ρ_z, ρ_ω, ρ_γ, ρ_τ and ρ_η, *i.e.*, we set $p(\mathbf{z}|\rho_z) = \prod_{i=1}^d \text{Bern}(z_i|\rho_z)$, $p(\boldsymbol{\omega}|\rho_\omega) = \prod_{k=1}^K \text{Bern}(\omega_k|\rho_\omega)$, $p(\boldsymbol{\gamma}|\rho_\gamma) = \prod_{i=1}^d \text{Bern}(\gamma_k|\rho_\gamma)$, $p(\{\boldsymbol{\tau}^{(k)}\}_{k=1}^K|\rho_\tau) = \prod_{k=1}^K\prod_{i=1}^d \text{Bern}(\tau_i^{(k)}|\rho_\tau)$ and finally $p(\{\boldsymbol{\eta}^{(k)}\}_{k=1}^K|\rho_\eta) = \prod_{k=1}^K\prod_{i=1}^d \text{Bern}(\eta_i^{(k)}|\rho_\eta)$. The hyper-prior for each ρ_z, ρ_ω, ρ_γ, ρ_τ and ρ_η is a beta distribution with parameters a_0 and b_0, *e.g.*, $p(\rho_z) = \text{Beta}(\rho_z|a_0, b_0)$ for the case of ρ_z. Furthermore, we set $a_0 = 1$ and $b_0 = 1$ which leads to a uniform distribution so that no particular hyper-parameter value is favored *a priori*. These uniform priors allow to identify each hyper-parameter value from the training data.

Last, we set the hyper-prior for the noise of each task to be an inverse gamma distribution, *i.e.*, $p(\boldsymbol{\sigma}^2) = \prod_{k=1}^K \text{InvGam}(\sigma^2_{(k)}|\alpha_0, \beta_0)$, where we specify $\alpha_0 = 5$ and $\beta_0 = 5$. These parameter values are equivalent to the prior observation in each task of 10 data instances with noise variance equal to 1. Furthermore, they also produce high variance in the prior distribution which allows for the identification of the correct level of noise of each task using the training data only. An alternative formulation of the prior that assumes the same level of noise for each task is also considered. Namely, we set $p(\boldsymbol{\sigma}^2) = \text{InvGam}(\sigma^2|\alpha_0, \beta_0)$, where each entry in $\boldsymbol{\sigma}^2$ is constrained to be equal to σ^2.

2.2. Prediction and identification of relevant features

Define $\boldsymbol{\rho} = \{\rho_z, \rho_\omega, \rho_\gamma, \rho_\tau, \rho_\eta\}$. The joint probability of the targets $\mathbf{Y}$ and the latent variables $\mathbf{W}$, $\mathbf{\Omega}$, $\boldsymbol{\rho}$ and $\boldsymbol{\sigma}^2$ is:

$$p(\mathbf{Y}, \mathbf{W}, \mathbf{\Omega}, \boldsymbol{\rho}, \boldsymbol{\sigma}^2|\boldsymbol{\mathcal{X}}) = p(\mathbf{Y}|\boldsymbol{\mathcal{X}}, \mathbf{W}, \boldsymbol{\sigma}^2)p(\mathbf{W}|\mathbf{\Omega})\times \\ \times\, p(\mathbf{\Omega}|\boldsymbol{\rho})p(\boldsymbol{\rho})p(\boldsymbol{\sigma}^2)\,, \tag{4}$$

where $p(\mathbf{Y}|\boldsymbol{\mathcal{X}}, \mathbf{W}, \boldsymbol{\sigma}^2)$ is given by (1), $p(\mathbf{W}|\boldsymbol{\Omega})$ is given by (3), $p(\boldsymbol{\Omega}|\boldsymbol{\rho}) = p(\mathbf{z}|\rho_z)p(\boldsymbol{\omega}|\rho_\omega)p(\boldsymbol{\gamma}|\rho_\gamma)$ $p(\{\boldsymbol{\tau}^{(k)}\}_{k=1}^K|\rho_\tau)p(\{\boldsymbol{\eta}^{(k)}\}_{k=1}^K|\rho_\eta)$ and $p(\boldsymbol{\rho}) = p(\rho_z)$ $p(\rho_\omega)p(\rho_\gamma)p(\rho_\tau)p(\rho_\eta)$. This joint distribution is normalized with respect to $\mathbf{W}$, $\boldsymbol{\Omega}$, $\boldsymbol{\rho}$ and $\boldsymbol{\sigma}^2$ to get a posterior:

$$p(\mathbf{W}, \boldsymbol{\Omega}, \boldsymbol{\rho}, \boldsymbol{\sigma}^2|\mathbf{Y}, \boldsymbol{\mathcal{X}}) = \frac{p(\mathbf{Y}, \mathbf{W}, \boldsymbol{\Omega}, \boldsymbol{\rho}, \boldsymbol{\sigma}^2|\boldsymbol{\mathcal{X}})}{p(\mathbf{Y}|\boldsymbol{\mathcal{X}})}. \quad (5)$$

The posterior is used to compute predictions for the target value y_{new} of a new un-observed instance $\mathbf{x}_{\text{new}}$ of task k. Namely, $p(y_{\text{new}}|\mathbf{x}_{\text{new}}) = \sum_{\boldsymbol{\Omega}} \int \mathcal{N}(y_{\text{new}}|\mathbf{x}_{\text{new}}^{\text{T}}\mathbf{w}^{(k)}, \sigma_{(k)}^2)p(\mathbf{W}, \boldsymbol{\Omega}, \boldsymbol{\rho}, \boldsymbol{\sigma}^2|\mathbf{Y}, \boldsymbol{\mathcal{X}})d\mathbf{W}$ $d\boldsymbol{\rho}d\boldsymbol{\sigma}^2$. The probability that a particular $w_i^{(k)}$ is different from zero can be computed similarly. For this, we eliminate variables in (5) and sum the posterior probabilities of the three events described in Section 2.1.2, *i.e.*, $p(w_i^{(k)} \neq 0|\mathbf{Y}, \boldsymbol{\mathcal{X}}) = p(\{z_i = 1 \cap \eta_i^{(k)} = 1\} \cup \{z_i = 0 \cap \omega_k = 1 \cap \tau_i^{(k)} = 1\} \cup \{z_i = 0 \cap \omega_k = 0 \cap \gamma_i = 1\}|\mathbf{Y}, \boldsymbol{\mathcal{X}})$. Finally, the probability that task k is an outlier, $p(\omega_k = 1|\mathbf{Y}, \boldsymbol{\mathcal{X}})$, or the probability that feature i is an outlier, $p(z_i = 1|\mathbf{Y}, \boldsymbol{\mathcal{X}})$, are computed in a similar way.

The computation of all the expressions described in this section, except (4), is intractable for typical problems. Thus, we have to resort to approximate inference methods.

3. Expectation propagation (EP)

EP is an efficient mechanism for approximate inference (Minka, 2001). EP approximates each factor in (4) that is not inside a particular exponential family $\mathcal{F}$ of distributions with an un-normalized factor that is inside $\mathcal{F}$. We set $\mathcal{F}$ to be the product of Gaussian distributions on $\mathbf{W}$, Bernoulli distributions on $\boldsymbol{\Omega}$, beta distributions on $\boldsymbol{\rho}$ and inverse gamma distributions on $\boldsymbol{\sigma}^2$. $\mathcal{F}$ is closed under product and division operations. The only factors in (4) that are not in $\mathcal{F}$ are those of the likelihood (1), $p(\mathbf{W}|\boldsymbol{\Omega})$ and $p(\boldsymbol{\Omega}|\boldsymbol{\rho})$. The hyper-prior for $\boldsymbol{\rho}$ is beta, and the hyper-prior for $\boldsymbol{\sigma}^2$ is inverse gamma so they need not be approximated.

In EP each likelihood factor corresponding to the n-th instance of the k-th task $(\mathbf{x}_n^{(k)}, y_n^{(k)})$ is approximated as $p(y_n^{(k)}|\mathbf{w}^{(k)}, \mathbf{x}_n^{(k)}, \sigma_{(k)}^2) = \mathcal{N}(y_n^{(k)}|(\mathbf{x}_n^{(k)})^{\text{T}}\mathbf{w}^{(k)}, \sigma_{(k)}^2) \approx \tilde{f}_n^{(k)}(\mathbf{w}^{(k)}, \sigma_{(k)}^2) = \tilde{c}_n^{(k)} \mathcal{N}((\mathbf{x}_n^{(k)})^{\text{T}}\mathbf{w}^{(k)}|\tilde{m}_n^{(k)}, \tilde{v}_n^{(k)})\text{InvGam}(\sigma_{(k)}^2|\tilde{a}_n^{(k)}, \tilde{b}_n^{(k)})$.

The approximation of each factor $p(w_i^{(k)}|\boldsymbol{\Omega})$ that appears in $p(\mathbf{W}|\boldsymbol{\Omega})$ in (3) is $p(w_i^{(k)}|\boldsymbol{\Omega}) \approx \tilde{g}_i^{(k)}(w_i, z_i, \omega_k, \gamma_i, \tau_i^{(k)}, \eta_i^{(k)}) = \tilde{s}_i^{(k)}\mathcal{N}(w_i|\tilde{m}_i^{(k)}, \tilde{\sigma}_{(i,k)}^2)$ $\text{Bern}(z_i|\tilde{p}_z^{(i,k)})\text{Bern}(\omega_k|\tilde{p}_\omega^{(i,k)})\text{Bern}(\gamma_i|\tilde{p}_\gamma^{(i,k)})\text{Bern}(\tau_i^{(k)}|\tilde{p}_\tau^{(i,k)})\text{Bern}(\eta_i^{(k)}|\tilde{p}_\eta^{(i,k)})$. Finally, each factor in $p(\boldsymbol{\Omega}|\boldsymbol{\rho})$ is approximated following a similar principle. For example, for $p(z_i|\rho_z)$ the approximation is $\text{Bern}(z_i|\rho_z) \approx \tilde{h}_z^{(i)}(z_i, \rho_z) = \tilde{\kappa}_z^{(i)}\text{Bern}(z_i|\tilde{p}_z^{(i)})\text{Beta}(\rho_z|\tilde{a}_z^{(i)}, \tilde{b}_z^{(i)})$. The approximation of the other factors in $p(\boldsymbol{\Omega}|\boldsymbol{\rho})$ is equivalent to this one. All the parameters with the superscript $\tilde{}$ are adjusted by EP, as described below.

The EP approximation of the joint distribution (4) replaces each exact factor by the corresponding approximate one. Denote by $\tilde{q}$ this approximation. After normalization, the joint distribution (4) becomes the exact posterior (5). Similarly, after normalization $\tilde{q}$ becomes the EP posterior approximation q:

$$q(\mathbf{W}, \boldsymbol{\Omega}, \boldsymbol{\rho}, \boldsymbol{\sigma}^2) = \frac{\tilde{q}(\mathbf{W}, \boldsymbol{\Omega}, \boldsymbol{\rho}, \boldsymbol{\sigma}^2)}{Z_q}, \quad (6)$$

which is inside of $\mathcal{F}$ because $\mathcal{F}$ is closed under the product operation. The parameters of q are obtained from the product of all the factors in $\tilde{q}$ and Z_q can be readily computed because $\tilde{q}$ is an un-normalized parametric distribution inside of $\mathcal{F}$. Given q, all the expressions in Section 2.2 can be approximated by replacing the exact posterior with q.

EP refines until convergence each approximate factor $\tilde{f}$. For this, $q^{\text{old}} \propto q/\tilde{f}$ is computed. q^{old} has the same form as q because $q, \tilde{f} \in \mathcal{F}$. Then, an updated posterior approximation q^{new} is obtained by minimizing the Kullback-Leibler divergence between fq^{old} and q^{new}, $\text{KL}(fq^{\text{old}}||q^{\text{new}})$, where f denotes the exact factor associated to $\tilde{f}$. The updated approximate factor is $\tilde{f} = Z_f q^{\text{new}}/q^{\text{old}}$, where Z_f is the normalization constant of fq^{old}. This guarantees that $\tilde{f}$ is similar to the exact factor in regions of high posterior probability, as estimated by q^{old}. The minimization of $\text{KL}(fq^{\text{old}}||q^{\text{new}})$ with respect to q^{new} has a global optimum found by matching expected sufficient statistics between fq^{old} and q^{new} (Bishop, 2006). These expectations can be obtained from the derivatives of $\log Z_f$ with respect to the (natural) parameters of q^{old}, as indicated by Seeger (2006).

A contribution of this paper is the computation of Z_f for the factors in $p(\mathbf{W}|\boldsymbol{\Omega})$. In that case, Z_f is a probabilistic mixture of the convolution of the Strawderman-Berger prior $\pi(\cdot)$ with a Gaussian distribution, *i.e.*, the posterior distribution for $w_i^{(k)}$ under q^{old}, and the convolution of the point probability mass at the origin, δ_0, with the same Gaussian. Fortunately, the Strawderman-Berger prior has a closed form convolution with the Gaussian distribution. Johnstone & Silverman (2005) provide the analytic solution when the variance of the Gaussian is one. We provide the solution when this is not the case. The complete details about EP are found in the supplementary material, alongside with an R implementation of the proposed method.

When $d > N_k$, where N_k is the number of instances of task k and d is the number of features, the covariance matrix of the likelihood of each task in (1) is low rank. EP is able to exploit this and it has a cost that scales like $\mathcal{O}(\sum_{k=1}^K N_k^2 d)$.

4. Related work

There are several works in the literature focusing on feature selection within a multi-task learning setting. In this section they are described. For example, Hernández-Lobato et al. (2010) propose a model based on the *spike-and-slab* prior (Mitchell & Beauchamp, 1988) to determine whether a feature is either relevant or irrelevant across all tasks. The *spike-and-slab* prior is also used in (Jebara, 2004), where a multi-task feature selection method is derived using the maximum entropy discrimination formalism. In (Obozinski et al., 2009; Vogt & Roth, 2010) the group LASSO is considered as an efficient estimator that selects common features across tasks by penalizing a mixed norm of the model coefficients. The work presented by Argyriou et al. (2007) is based on a similar approach. Finally, in (Xiong et al., 2007) a set of common relevant features across tasks is found using the automatic relevance determination principle. In summary, all these works assume jointly relevant and irrelevant features across tasks, as *e.g.* in Figure 2 (top), and are hence expected to perform poorly when this hypothesis is not fully satisfied, as *e.g.* in Figure 2 (bottom).

There are other methods that have been proposed to relax the hypothesis of jointly relevant and irrelevant features across all tasks. In (Jalali et al., 2010) a dirty model considers a mixed norm to penalize the model coefficients of several tasks. Specifically, $\mathbf{W} = \mathbf{P} + \mathbf{Q}$ where $\mathbf{P}$ is penalized with the ℓ_1 norm and $\mathbf{Q}$ with the the $\ell_{1,\infty}$ norm. A similar model can be derived using the $\ell_{1,2}$ norm instead (Vogt & Roth, 2010). The estimator employed selects a common subset of relevant features for all the tasks, but it allows for tasks with additional specific relevant features. The result is a generalization of the group LASSO (Obozinski et al., 2009; Vogt & Roth, 2010), which is expected to be more robust to outlier tasks in the learning process, but not to be as flexible as the model proposed in this paper. Another method introduced for this purpose is found in (Gong et al., 2012). This robust multi-task feature learning model also defines $\mathbf{W} = \mathbf{P} + \mathbf{Q}$ and estimates the model coefficients $\mathbf{W}$ by penalizing both $\mathbf{P}$ and $\mathbf{Q}^{\mathrm{T}}$ with the $\ell_{1,2}$ norm. The intuition behind this idea is that if the k-th row of $\mathbf{Q}$ is not zero after the estimation, task k is an outlier task with all features relevant for prediction. However, again this a less flexible assumption than the one we make. In particular, the two works just described assume that all tasks share a few relevant features, although they allow for some arbitrary tasks to have extra relevant features. This means that they cannot model outlier tasks, (*e.g.*, tasks 4 and 8 in Figure 2 (bottom)), unlike the approach proposed in this work.

Hernández-Lobato & Hernández-Lobato (2013) consider that tasks do not share common relevant and irrelevant features for prediction, but common dependencies in the feature selection process. These dependencies are induced from the data using a generalization of the horseshoe prior for feature selection (Carvalho et al., 2009). The principle they follow is hence more flexible than the assumption made by the models described in the first paragraph of this section. However, such an approach is expected to be suboptimal when most tasks actually share relevant and irrelevant features for prediction, which is the hypothesis we assume and the one that is displayed in Figure 2 (bottom).

Finally, some works in the literature also consider modeling outlier tasks in multi-task learning, *e.g.*, (Xue et al., 2007; Passos et al., 2012). However, they do not consider sparsity in the model coefficients and are hence expected to perform poorly when this hypothesis is actually satisfied in practice.

5. Experiments

We compare the proposed model for dirty multi-task feature selection (DMFS) with single task learning (STL) and with a model for multi-task feature selection (MFS) that assumes relevant and irrelevant features shared across all tasks. STL and MFS are particular cases of DMFS with all tasks being outliers (STL) and with no outlier tasks nor outlier features (MFS). We also compare results with the methods described in Section 4. That is, the dirty model (DM) of Jalali et al. (2010), the robust multi-task feature learning method (RMFL) of Gong et al. (2012) and the model for learning feature selection dependencies (MFS_{Dep}) of Hernández-Lobato & Hernández-Lobato (2013). In DM and RMFL we choose hyper-parameters using a grid search guided by an inner cross-validation method. In MFS_{Dep} we use type-II maximum likelihood for this (Bishop, 2006). DMFS, STL and MFS need not fix any hyper-parameters since they infer them from the data using hyper-priors. Unless stated differently, in all probabilistic models we assume different levels of noise for each task when training. All methods described are implemented in the R language.

5.1. Experiments with synthetic data

We generate 12 tasks where the model coefficients are sampled from a Student's distribution with 5 degrees of freedom. Each task k has $d = 2000$ attributes and $N_k = 150$ samples. The sparsity pattern employed for the model coefficients across tasks is displayed in Figure 2 (bottom). All model coefficients above dimension 26 are equal to zero. The targets are added Gaussian noise with variance $1/2$ and each entry of the design matrix $\mathbf{X}^{(k)}$ of task k is standard Gaussian. We use 90% of the instances for training and 10% for testing. The reported estimates are averaged over 100 repetitions. We report the test root mean squared error (RMSE) and the average reconstruction error of the model coefficients, *i.e.*, $1/K \sum_{k=1}^{K} ||\mathbf{w}^{(k)} - \hat{\mathbf{w}}^{(k)}||_2$, where $\hat{\mathbf{w}}^{(k)}$ is either the posterior mean (only in the probabilistic models), or a point estimate of $\mathbf{w}^{(k)}$ (only in DM and RMFL).

Table 1. Avg. test RMSE, reconstruction error and running time in minutes of each method on the synthetic experiments.

Method	Test RMSE	Rec. Error	Training Time
MFS	0.81±0.06	0.37±0.04	6.41±1.57
DMFS	**0.73 ±0.04**	**0.22±0.02**	21.87±0.18
DM	0.86±0.05	0.50±0.03	150.35±10.0
MFS_{DEP}	0.77±0.06	0.32±0.04	$2 \cdot 10^3 \pm 4 \cdot 10^2$
RMFL	0.90±0.05	0.56±0.03	95.42±5.0
STL	0.78±0.08	0.33±0.06	**5.14±0.39**

The results obtained are displayed in Table 1. The best method in terms of the reconstruction error and the RMSE is DMFS, followed by MFS_{Dep} and STL. MFS performs worse than STL. DM and RMFL perform poorly. The differences of DMFS with respect to the other methods are statistically significant (p-value $< 5\%$ using a paired Student's T test). In terms of training time, the fastest method is STL closely followed by MFS and DMFS. DM and RMFL take longer training times due to the expensive grid search procedure that is used to fix their two hyper-parameters. This process demands re-training each method many times. If the optimal hyper-parameters were given, they would be the fastest methods. The training time of MFS_{Dep} is very high for the same reason. By contrast, unlike these methods, DMFS uses Bayesian inference to infer hyper-parameters and does not require any re-training.

The better results obtained by DMFS are also explained by Figure 3, which shows the average posterior probability that each task and each feature is an outlier, as estimated by DMFS. DMFS successfully identifies tasks 4 and 8 as outlier tasks and features 19 and 21 as outlier features. We note that features 1, 3 and 24 have also a small probability of being outliers. This makes sense because according to Figure 2 (bottom) they are relevant only for a few tasks.

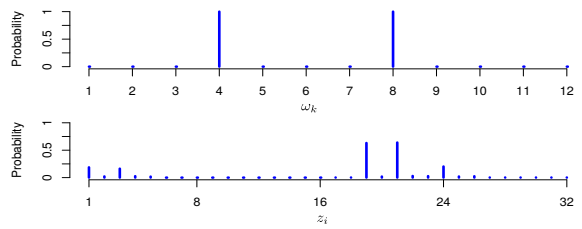

Figure 3. Avg. posterior probability for $\omega_k = 1$ (task k is an outlier) and $z_i = 1$ (feature i is an outlier) in DMFS, in the synthetic data. The last prob. is only displayed for the first 32 features.

Figure 4 also sheds light on the better performance of DMFS. It shows in a gray scale the average probability that each of the 26 first model coefficients of each task is zero, as estimated by each method. These probabilities are obtained from the approximate posterior in the probabilistic models. In DM and RMFL we report the fraction of times that a coefficient is different from zero across experiments. We note that DMFS, MFS_{Dep} and STL find patterns that agree the most with those of Figure 2 (bottom). However, STL does not exploit the multi-task setting and is less confident about the non-zeroness of coefficients 6 to 16 for non-outlier tasks. DMFS is also more confident than MFS_{Dep} about irrelevant coefficients. On the other hand, MFS, RMFL and DM give high probability of being different from zero to coefficients 6 to 16 for tasks 4 and 8. The reason is that they assume a few relevant features shared across all tasks and cannot model outlier tasks. Furthermore, they also find to be non-zero across all tasks some coefficients corresponding to features that are in fact only relevant for a few tasks (*e.g.*, the coefficients corresponding to features 1, 3, 19, 21 and 24). DM and RMFL can in principle model some of these coefficients (note that they give higher probabilities of being not zero to some of them), but to avoid their joint selection, these methods would have to shrink non-zero coefficients shared by all non-outlier tasks, producing worse results. In particular, the norms that they use cannot provide very sparse solutions and not shrink relevant coefficients (Hernández-Lobato et al., 2013).

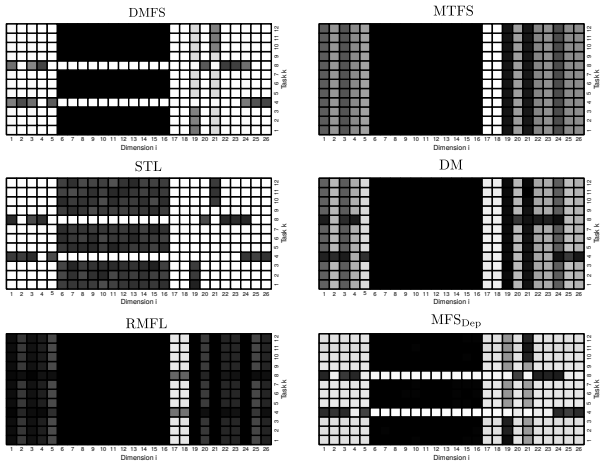

Figure 4. Average probability for each method across the 100 repetitions that each of the 26 first model coefficients of each task is different from zero in a gray scale (0 = white and 1 = black).

5.2. Reconstruction of gene regulatory networks

Assume $\mathbf{X}$ is a $N \times d$ matrix with columns denoting d genes and rows containing N measurements of log mRNA concentration obtained under different steady state conditions. Consider that $\mathbf{X}$ is contaminated with additive Gaussian noise. Then, $\mathbf{X} \approx \mathbf{X}\mathbf{W}^{\text{T}} + \sigma^2\mathbf{E}$, where the entries in $\mathbf{E}$ are standard Gaussian, σ^2 is the variance of the noise and $\mathbf{W}$ is a sparse $d \times d$ regression matrix with zero diagonal entries that links the expression level of each gene with that of its transcriptional regulators (Hernández-Lobato et al., 2015). When an entry of $\mathbf{W}$ is non-zero there is a regulatory dependency between the pair of genes it refers to. These dependencies are described by gene regulatory networks in which there is a node per gene and two nodes are connected with a directed edge if the first gene regulates

the second. These networks are sparse (with many missing edges) and have hub nodes (transcription factors) that regulate several genes. Figure 5 shows a sample network.

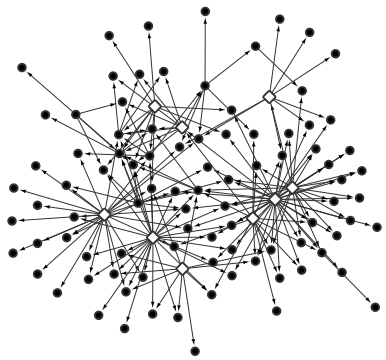

Figure 5. Sample gene regulatory network used in the experiments. Nodes that are potential hubs have a diamond shape.

Table 2. Avg. area under the ROC curve for the network reconstruction experiments and RMSE for the anticancer drug sensitivity experiments, for each of the different methods considered.

Method	AUROC	RMSE
MFS	0.73±0.05	0.733±0.053
DMFS	**0.84±0.05**	0.717±0.050
DM	0.76±0.06	**0.703±0.050**
MFS_{DEP}	0.79±0.05	0.704±0.051
RMFL	0.79±0.05	**0.703±0.050**
STL	0.70±0.04	0.730±0.049

We formulate the problem of inducing $\mathbf{W}$ given $\mathbf{X}$ as a multi-task problem with d tasks where the model coefficients of task k correspond to the k-th row of $\mathbf{W}$. The design matrix $\mathbf{X}^{(k)}$ is given by the matrix $\mathbf{X}$ with column k set to zero. The targets of task k are the entries in the k-th column of $\mathbf{X}$. To favor sparse networks we use the proposed prior for $\mathbf{W}$. This prior models the hub nodes in the network by considering jointly relevant features across tasks, but it allows for small deviations to consider genes regulated, in addition, by a few extra genes (outlier features), or genes regulated by very specific transcription factors (outlier tasks). The regulatory network can be induced by computing the posterior probability p_{ij} that each entry $w_i^{(j)}$ in $\mathbf{W}$ is non-zero. The corresponding directed edge $j \rightarrow i$ is predicted when p_{ij} exceeds a threshold $\zeta \in [0, 1]$. Thus, these experiments evaluate the ability of each method to discriminate between relevant and irrelevant coefficients.

We evaluate the different methods in the task of inferring gene regulatory networks. The experimental protocol follows the DREAM 4 *in silico* challenge 2009. We generate 100 networks with 100 genes and sample 90 steady-state measurements from each network using GeneNetWeaver (Schaffter et al., 2011). The reconstruction performance is measured in terms of the area under the ROC curve (AUROC) (Fawcett, 2006), obtained when ζ varies between 0 and 1. In MFS, DM and RMFL, to induce the network, we use the absolute values of the estimated entries of $\mathbf{W}$ normalized to sum to one, instead of posterior probabilities.

Table 2 shows the average AUROC for each method. The best method (higher is better) is DMFS followed by MFS_{Dep}, RMFL, DM and MFS. The method with the lowest performance is STL. The differences are statistically significant (p-value $< 5\%$ using a paired Student's T test). This result shows that multi-task methods are beneficial in this problem and that the hypothesis made by DMFS is more adequate, probably because it is more flexible. In DMFS several tasks have a significantly higher probability of being outlier tasks, and the same is observed for several features (results not shown). We have also evaluated here the winning solution of the DREAM 4 challenge (Huynh-Thu et al., 2010). This method uses tree-ensembles to identify relevant features, but does not exploit task relations. The average AUROC obtained is 0.75, which is below the one shown in Table 2 for DMFS, MFS_{Dep} and RMFL.

5.3. Denoising of natural images

We consider the problem of denoising the 256×256 *house* image used in (Titsias & Lázaro-Gredilla, 2011) when it has been contaminated by Gaussian noise. Three levels are considered for the standard deviation of the noise $\sigma_{(k)}$. Namely, 25, 50 and 75, $\forall k$. Following that work, we partition the noisy image in $62,001$ overlapping blocks of 8×8 pixels and regard each block as a different task. These tasks are then grouped forming 64 groups of non-overlapping blocks (*i.e.*, one group of 32×32 blocks, 7 groups of 32×31 blocks, 7 groups of 31×32 blocks and 49 groups of 31×31 blocks) which are solved in parallel in a cluster using each multi-task method (see the supplementary material). To denoise the image we set $\mathbf{y}^{(k)}$ equal to each block and each $\mathbf{X}^{(k)}$ equal to an orthonormal basis corresponding to the *Haar* wavelet. Thus, $d = 64$ and $N_k = 64$ for each task k. It is well known that natural images have sparse representations under a wavelet basis. Thus, the learning process involves finding the wavelet coefficients corresponding to each block from the noisy observations. Using these coefficients the original image can be reconstructed by obtaining their projection under the wavelet basis. We assume in all probabilistic methods the same level of noise for each task.

Table 3 shows the peak-signal-to-noise ratio obtained by each method (higher is better) in the denoising process. The best results are obtained by the proposed approach DMFS, which improves the results of the other methods, especially for high levels of noise, where multi-task methods show a clear advantage over single-task learning. As in the previous experiments, DMFS also identifies here several outlier tasks and features (results not shown). Figure 6 shows the original noisy images and the corresponding denoised images obtained by DMFS for each value of $\sigma_{(k)}$.

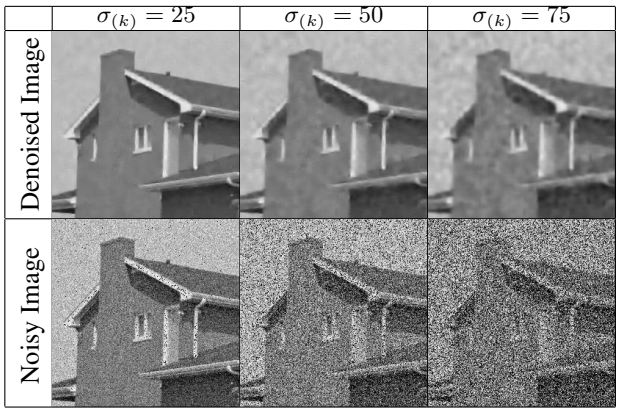

Figure 6. Noisy images and corresponding denoised images obtained by DMFS, for each different value of $\sigma_{(k)}$ considered $\forall k$.

Table 3. Peak-signal-to-noise ratio for each method.

Method	$\sigma_{(k)} = 25$	$\sigma_{(k)} = 50$	$\sigma_{(k)} = 75$
MFS	25.90	23.89	23.87
DMFS	**30.67**	**27.25**	**25.22**
DM	28.50	25.91	24.24
MFS_{DEP}	30.46	25.74	23.65
RMFL	28.35	25.56	24.09
STL	30.58	26.37	23.35

5.4. Anticancer drug sensitivity prediction

We consider the dataset described in (Barretina *et al.*, 2012). This dataset contains microarray gene expression data from 479 human cancer cell lines with pharmacological profiles for 24 anticancer drugs. After removing missing values 294 cell lines remain. We filter the data and consider only the $1,000$ genes with the largest interquartile distance. The task of interest is to predict each drug sensitivity (measured in terms of the area over the dose-response curve) from the microarray data for each cell line. Thus, in these experiments $d = 1,000$, $K = 24$ and $N_k = 294$ $\forall k$. We use 90% of the data for training and 10% for testing. The reported estimates are averaged over 100 repetitions. We report the test root mean squared error (RMSE). In these experiments assuming in all probabilistic methods the same level of noise for each task also improves results.

Table 2 shows the results of the experiments. The best methods are DM, RMFL and MFS_{Dep}. The solution of DM reduces to the one of the group LASSO (*i.e.*, one regularization parameters is set always to zero). We believe the better performance obtained by DM and RMFL is a consequence of shrinking too much relevant coefficients, which may be useful here to alleviate over-fitting since microarray data is notoriously very noisy. DMFS performs worse than these three methods, and the differences are statistically significant according to a paired Wilcoxon test (p-value $< 5\%$). Nevertheless, DMFS improves over the baselines STL and MFS, and the differences are also statistically significant. Finally, Figure 7 shows the average probability that each drug is an outlier task, as estimated by DMFS. We observe that several drugs are marked as outlier tasks.

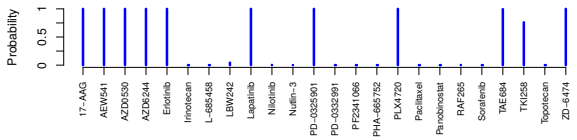

Figure 7. Avg. posterior probability that each drug is an outlier task, as estimated by DMFS, in the drug sensitivity experiments.

Last, we compare here the utility of DMFS to identify outlier tasks with that of RMFL. For this, we train the group LASSO on the data when the tasks identified as outliers by each method are removed (recall that the group LASSO performs best). In DMFS we remove those tasks whose probability of being an outlier is above 1%. In RMFL we remove those tasks whose rows in $\mathbf{Q}$ are not zero. The results obtained indicate that when DMFS is used to remove outlier tasks the RMSE of the group LASSO on the non-outlier tasks is 0.667 ± 0.061, when the outlier tasks are removed, and 0.671 ± 0.059 otherwise. This improvement is statistically significant according to a paired Wilcoxon test (p-value $< 5\%$). By contrast, when RMFL is used to remove outlier tasks the RMSE of the group LASSO on the non-outlier tasks is 0.684 ± 0.074, when the outlier tasks are removed, and 0.686 ± 0.069 otherwise. This other improvement is not statistically significant (p-value $> 5\%$), which shows that DMFS is better for identifying outlier tasks.

6. Conclusions

Most methods for multi-task feature selection in the literature assume jointly relevant and irrelevant features across tasks. This hypothesis may be too restrictive in practice. In this paper, we have proposed a new prior distribution that considers that most tasks share relevant and irrelevant features, but that allows for some tasks to have different relevant and irrelevant coefficients (outlier tasks), and for some features to be arbitrarily relevant or irrelevant for each task (outlier features). This is a more flexible assumption. Unfortunately, exact inference is infeasible under such a prior. Nevertheless, a quadrature-free expectation propagation method can be used for approximate inference. A model using our prior has been evaluated in several experiments involving the reconstruction of gene regulatory networks, the denoising of natural images and the prediction of drug sensitivity from microarray data. These experiments show gains in the prediction performance and in the identification of relevant features. Such a prior is also useful to better understand the data, since it allows to identify outlier tasks and features. When outlier tasks are removed from the training set, traditional multi-task feature selection methods obtain better results in the non-outlier tasks. This confirms that removed tasks were indeed outlier tasks.

Acknowledgements

Daniel Hernández-Lobato gratefully acknowledges the use of the facilities of Centro de Computacin Cientfica (CCC) at Universidad Autónoma de Madrid. This author also acknowledges financial support from Spanish Plan Nacional I+D+i, Grant TIN2013-42351-P, and from Comunidad de Madrid, Grant S2013/ICE-2845 CASI-CAM-CM. José Miguel Hernández-Lobato acknowledges financial support from the Rafael del Pino Fundation.

References

Argyriou, A., Evgeniou, T., and Pontil, M. Multi-task feature learning. In *Neural Information Processing Systems*, pp. 41–48. 2007.

Armagan, A., Dunson, D., and Clyde, M. Generalized beta mixtures of Gaussians. In *Neural Information Processing Systems*, pp. 523–531. 2011.

Barretina *et al*. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483: 603–307, 2012.

Berger, J. A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *The Annals of Statistics*, 8:716–761, 1980.

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006.

Carvalho, C.M., Polson, N.G., and Scott, J.G. Handling sparsity via the horseshoe. *J. Mach. Learn. Res. W&CP*, 5:73–80, 2009.

Fawcett, T. An introduction to ROC analysis. *Pattern recognition letters*, 27:861–874, 2006.

Gong, P., Ye, J., and Zhang, C. Robust multi-task feature learning. In *International Conference on Knowledge Discovery and Data Mining*, pp. 895–903, 2012.

Hernández-Lobato, D. and Hernández-Lobato, J. M. Learning feature selection dependencies in multi-task learning. In *Neural Information Processing Systems*, pp. 746–754. 2013.

Hernández-Lobato, D., Hernández-Lobato, J. M., Helleputte, T., and Dupont, P. Expectation propagation for Bayesian multi-task feature selection. In *European Conference on Machine Learning*, pp. 522–537, 2010.

Hernández-Lobato, D., Hernández-Lobato, J. M., and Dupont, P. Generalized spike-and-slab priors for Bayesian group feature selection using expectation propagation. *J. Mach. Learn. Res.*, 14:1891–1945, 2013.

Hernández-Lobato, J. M., Hernández-Lobato, D., and Suárez, A. Expectation propagation in linear regression models with spike-and-slab priors. *Machine Learning*, 99:437–487, 2015.

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5:e12776, 2010.

Jalali, A., Ravikumar, P., Sanghavi, S., and Ruan, C. A dirty model for multi-task learning. In *Neural Information Processing Systems*, pp. 964–972. 2010.

Jebara, T. Multi-task feature and kernel selection for SVMs. In *International Conference on Machine Learning*, pp. 55–62, 2004.

Johnstone, I. M. and Silverman, B. W. Empirical Bayes selection of wavelet thresholds. *Annals of Statistics*, 33:1700–1752, 2005.

Minka, T. Expectation propagation for approximate Bayesian inference. In *Annual Conference on Uncertainty in Artificial Intelligence*, pp. 362–36, 2001.

Mitchell, T. J. and Beauchamp, J. J. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83:1023–1032, 1988.

Obozinski, G., Taskar, B., and Jordan, M.I. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, pp. 1–22, 2009.

Passos, A., Rai, P., Wainer, J., and Daumé III, H. Flexible modeling of latent task structures in multitask learning. In *International Conference on Machine Learning*, 2012.

Schaffter, T., Marbach, D., and Floreano, D. Genenetweaver: In silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27:2263–2270, 2011.

Seeger, M. Expectation propagation for exponential families. Technical report, UC, Berkeley, 2006.

Strawderman, W. E. Proper Bayes minimax estimators of the multivariate normal mean. *The Annals of Mathematical Statistics*, 42:385–388, 1971.

Titsias, M. and Lázaro-Gredilla, M. Spike and slab variational inference for multi-task and multiple kernel learning. In *Neural Information Processing Systems*, pp. 2339–2347, 2011.

Vogt, J. E. and Roth, V. The group-lasso: $\ell_{1,\infty}$ regularization versus $\ell_{1,2}$ regularization. In *32nd Anual Symposium of the German Association for Pattern Recognition*, volume 6376, pp. 252–261, 2010.

Xiong, T., Bi, J., Rao, B., and Cherkassky, V. Probabilistic joint feature selection for multi-task learning. In *Seventh SIAM International Conference on Data Mining*, pp. 332–342, 2007.

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

Zhang, J., Ghahramani, Z., and Yang, Y. Flexible latent variable models for multi-task learning. *Machine Learning*, 73:221–242, 2008.