A Probabilistic Model for Dirty Multi-task Feature Selection

Daniel Hernández-Lobato
Universidad Autónoma de Madrid
daniel.hernandez@uam.es

José Miguel Hernández-Lobato
Harvard University
jmh@seas.harvard.edu

Zoubin Ghahramani
Cambridge University
zoubin@eng.cam.ac.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. Our experiments show that a model based on the new robust prior obtains better predictive performance than other benchmark methods.

1 Introduction

Multi-task feature selection methods are used to improve the learning of model coefficients from the observed data under the sparsity assumption [1, 2, 3, 4, 5]. In these methods several learning tasks that have a common feature space are solved simultaneously. Furthermore, it is often assumed that the tasks share relevant and irrelevant features, as illustrated by Figure 2 (left). Unfortunately, in some situations this hypothesis may be too restrictive [6]. Figure 2 (right) shows this scenario, in which a few of the tasks may have specific relevant and irrelevant features (outlier tasks) and a few of the features may be arbitrarily relevant and irrelevant across tasks (outlier features). In this situation, traditional multi-task feature selection methods are expected to perform poorly. To deal with these situations, in this paper we propose a multi-task feature selection model that is expected to have better properties in the presence of diverse tasks, *i.e.*, data with the properties described above. The model is based on a robust prior distribution for enforcing sparsity in the model coefficients. Exact inference is intractable under this prior. However, expectation propagation can be used for efficient approximate inference [7]. Our experiments illustrate the benefits of the model proposed. Specifically, it has better prediction properties than other methods from the literature. This model can also be used to identify relevant attributes for prediction, and outlier tasks and outlier features.

2 Model Description

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 of targets for task k, respectively. All tasks share the same d attributes or features. 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 $\boldsymbol{\epsilon}^{(k)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma^2_{(k)})$. Let $\mathbf{W}$ be a $K \times d$ ma-

trix whose k-th row is $\mathbf{w}^{(k)}$ and $\mathbf{Y}$ a matrix whose k-th row is $\mathbf{y}^{(k)}$. The likelihood for $\mathbf{W}$ is $p(\mathbf{Y}|\{\mathbf{X}^{(k)}\}_{k=1}^K, \mathbf{W}, \{\sigma_{(k)}^2\}_{k=1}^K) = \prod_{k=1}^K \mathcal{N}(\mathbf{y}^{(k)}|\mathbf{X}^{(k)}\mathbf{w}^{(k)}, \mathbf{I}\sigma_{(k)}^2))$. Furthermore, feature selection for each task, or equivalently, sparsity in $\mathbf{w}^{(k)}$ is expected to be beneficial. We also assume that the K tasks share, in general, 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 $\mathbf{W}$.

2.1 Robust prior distribution

To favor sparse solutions we use the discrete mixture prior described in [8]. After reviewing this prior we extend it 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 [8] it is suggested for $\pi(\cdot)$ the Strawderman-Bergen prior [9, 10], which has Cauchy-like tails and yet allows for a closed form convolution with the Gaussian likelihood. This discrete mixture prior is a scale mixture of Gaussians:

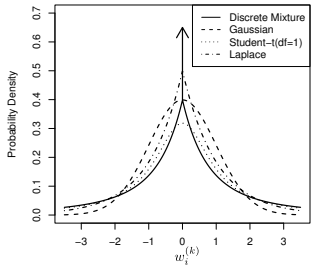

Figure 1: Density of different priors.

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

where $|\cdot|$ denotes absolute value, and $\Phi(\cdot)$ and $\mathcal{N}(\cdot|0,1)$ respectively denote the cdf and density of a standard Gaussian distribution. Figure 1 compares the discrete mixture prior with other priors (an arrow denotes 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. Thus, such a prior is very convenient for feature selection [8].

2.1.2 A robust prior to favor sparse solutions across tasks

The discrete mixture prior is extended to carry out feature selection across several tasks. We assume that these tasks have in general jointly relevant and irrelevant features. However, we consider a few outlier tasks with specific relevant and irrelevant features. Similarly, we also consider a few outlier features that may be arbitrary relevant and irrelevant for each task. This is illustrated in Figure 2 (right). Tasks 4 and 8 are outlier tasks and features 19 and 21 are outlier features. The remaining tasks and features follow the main assumption of jointly relevant and irrelevant features.

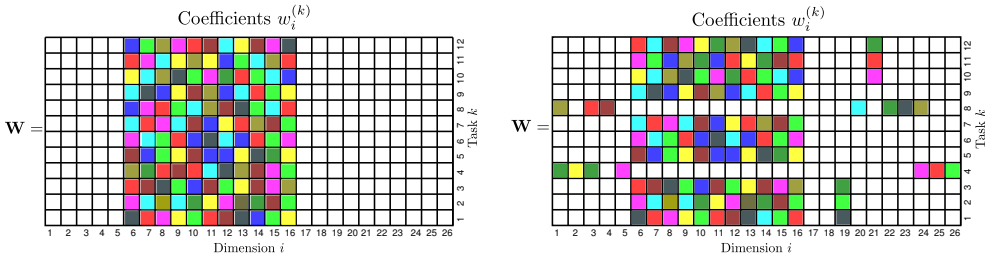

Figure 2: (left) Traditional multi-task feature selection: All tasks share relevant and irrelevant features (model coefficients). (right) Dirty multi-task feature selection: Most tasks share relevant and irrelevant features, but we allow for outlier tasks (tasks 4 and 8) and for outlier features (dimensions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

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

Consider $\mathbf{\Omega}$ to summarize all these latent variables, *i.e.* $\mathbf{\Omega} = \{\mathbf{z}, \boldsymbol{\omega}, \boldsymbol{\gamma}, \{\boldsymbol{\tau}^{(k)}\}_{k=1}^K, \{\boldsymbol{\eta}^{(k)}\}_{k=1}^K\}$. We specify the prior distribution for the model coefficients to be $p(\mathbf{W}|\mathbf{\Omega}) = \prod_{i=1}^d \prod_{k=1}^K p(w_i^{(k)}|\mathbf{\Omega})$, 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 ($z_i = 1$) which is 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.

We fix the hyper-priors for the latent variables to Bernoullis with parameters ρ_z, ρ_ω, ρ_γ, ρ_τ and ρ_η. That is, $p(\mathbf{z}) = \prod_{i=1}^d \text{Bern}(z_i|\rho_z)$, $p(\boldsymbol{\omega}) = \prod_{k=1}^K \text{Bern}(\omega_k|\rho_\omega)$, $p(\boldsymbol{\gamma}) = \prod_{i=1}^d \text{Bern}(\gamma_k|\rho_\gamma)$, $p(\{\boldsymbol{\tau}^{(k)}\}_{k=1}^K) = \prod_{k=1}^K \prod_{i=1}^d \text{Bern}(\tau_i^{(k)}|\rho_\tau)$ and $p(\{\boldsymbol{\eta}^{(k)}\}_{k=1}^K) = \prod_{k=1}^K \prod_{i=1}^d \text{Bern}(\eta_i^{(k)}|\rho_\eta)$. 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. We set $a_0 = 1$ and $b_0 = 10$. These values favor sparse solutions (hyper-parameter values close to zero) and, at the same time, produce high variance to identify of the correct hyper-parameter values from the training data.

3 Expectation Propagation

Define $\boldsymbol{\rho} = \{\rho_z, \rho_\omega, \rho_\gamma, \rho_\tau, \rho_\eta\}$ and $p(\boldsymbol{\rho}) = \prod_{\rho \in \boldsymbol{\rho}} p(\rho)$. The joint distribution $p(\mathbf{Y}, \mathbf{W}, \mathbf{\Omega}, \boldsymbol{\rho}|\{\mathbf{X}^{(k)}\}_{k=1}^K, \{\sigma_{(k)}^2\}_{k=1}^K) = p(\mathbf{Y}|\{\mathbf{X}^{(k)}\}_{k=1}^K, \mathbf{W}, \{\sigma_{(k)}^2\}_{k=1}^K) p(\mathbf{W}|\mathbf{\Omega}) p(\mathbf{\Omega}|\boldsymbol{\rho}) p(\boldsymbol{\rho})$ can be normalized with respect to $\mathbf{W}$, $\mathbf{\Omega}$ and $\boldsymbol{\rho}$ to get a posterior distribution over the latent variables,

$$p(\mathbf{W}, \mathbf{\Omega}, \boldsymbol{\rho}|\mathbf{Y}, \{\mathbf{X}^{(k)}\}_{k=1}^K, \{\sigma_{(k)}^2\}_{k=1}^K) = \frac{p(\mathbf{Y}, \mathbf{W}, \mathbf{\Omega}, \boldsymbol{\rho}|\{\mathbf{X}^{(k)}\}_{k=1}^K, \{\sigma_{(k)}^2\}_{k=1}^K)}{p(\mathbf{Y}|\{\sigma_{(k)}^2\}_{k=1}^K)}, \quad (2)$$

whose exact computation is intractable in most real applications. To circumvent this problem, EP approximates (2) by replacing each factor in $p(\mathbf{Y}, \mathbf{W}, \mathbf{\Omega}, \boldsymbol{\rho}|\{\mathbf{X}^{(k)}\}_{k=1}^K, \{\sigma_{(k)}^2\}_{k=1}^K)$ that is not inside a particular exponential family $\mathcal{F}$ of distributions with an approximate factor inside that particular family [7]. We set $\mathcal{F}$ to be the product of Gaussian distributions on $\mathbf{W}$, Bernoulli distributions on $\mathbf{\Omega}$ and beta distributions on $\boldsymbol{\rho}$. Therefore, the only factors not in $\mathcal{F}$ are $p(\mathbf{W}|\mathbf{\Omega})$ and $p(\mathbf{\Omega}|\boldsymbol{\rho})$. The likelihood is Gaussian and the hyper-priors are beta. Thus, these factors need not be approximated.

In our EP method each factor $p(w_i^{(k)}|\mathbf{\Omega})$ in $p(\mathbf{W}|\mathbf{\Omega})$ is approximated as $p(w_i^{(k)}|\mathbf{\Omega}) \approx \tilde{s}_i^{(k)} \mathcal{N}(w_i|\tilde{m}_i^{(k)}, \tilde{\sigma}_{(i,k)}^2) \text{Bern}(z_i|\tilde{p}_z^{(i,k)}) \text{Bern}(\omega_k|\tilde{p}_\omega^{(i,k)}) \text{Bern}(\gamma_i|\tilde{p}_\gamma^{(i,k)}) \text{Bern}(\tau_i^{(k)}|\tilde{p}_\tau^{(i,k)}) \text{Bern}(\eta_i^{(k)}|\tilde{p}_\eta^{(i,k)})$. Each factor in $p(\mathbf{\Omega}|\boldsymbol{\rho})$ is approximated similarly. Namely, for the particular case of ρ_z, $\text{Bern}(z_i|\rho_z) \approx \tilde{\kappa}_z^i \text{Bern}(z_i|\tilde{p}_z^{(i)}) \text{Beta}(\rho_z|\tilde{a}_z^{(i)}, \tilde{b}_z^{(i)})$. Furthermore, all the parameters with the superscript $\tilde{}$ are to be adjusted by EP. EP does this so that the approximate factors look similar to the corresponding exact factors in regions of high posterior probability. Once this fitting process is finished, the EP approximation of (2) is obtained by replacing in the joint distribution $p(\mathbf{Y}, \mathbf{W}, \mathbf{\Omega}, \boldsymbol{\rho}|\{\mathbf{X}^{(k)}\}_{k=1}^K, \{\sigma_{(k)}^2\}_{k=1}^K)$ each exact factor by the corresponding approximate factor. Denote with $\tilde{q}$ the resulting approximate joint distribution. After normalization, $\tilde{q}$ becomes the EP posterior approximation q, which is inside of $\mathcal{F}$ because $\mathcal{F}$ is closed under the product operation.

4 Experimental evaluation

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 jointly relevant and irrelevant features across tasks. STL and MFS are particular cases of DMFS where all tasks are outliers (STL) and where there are no outlier tasks nor outlier features (MFS). We also compare with the dirty model (DM) described in [6], the robust multi-task feature selection method (RMFS) given in [11] and the model for learning feature selection dependencies (MFS_{Dep}) proposed in [12]. Besides these, other works from the literature also model outlier tasks, *e.g.*, [13, 14]. However, they do not consider sparsity in the model coefficients and are expected to perform poorly in our setting.

We generate 12 tasks where the model coefficients are sampled from a Student's distribution with 5 degrees of freedom. Each task k has $d = 200$ associated attributes and $n_k = 100$ samples. The sparsity pattern employed for the model coefficients is displayed in Figure 2 (right). All coefficients above dimension 26 are set to zero. The noise is Gaussian and $\sigma^2_{(k)} = 0.5\ \forall k$. Each entry of $\mathbf{X}^{(k)}$ is standard Gaussian $\forall k$. We use 90% of the instances for training and 10% for testing. We average results over 100 repetitions. For each method we report the test root mean squared error (RMSE) and the reconstruction error of $\mathbf{W}$, *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 (probabilistic models) or a point estimate of $\mathbf{w}^{(k)}$ (DM and RMFS). In the probabilistic models we set $\sigma^2_{(k)} = 1/2, \forall k$. In DM and RMFS we choose hyper-parameters using a grid of values and an inner cross-validation method. In MFS_{Dep} we use type-II maximum likelihood [15] for this purpose.

Table 1: Avg. Test RMSE and Reconstruction Error.

Method	Test RMSE	Rec. Error
MFS	0.80±0.06	0.41±0.03
DMFS	**0.74±0.05**	**0.27±0.02**
DM	0.89±0.06	0.58±0.04
MFS_{Dep}	0.76±0.05	0.32±0.02
RMFS	0.95±0.08	0.65±0.06
STL	0.78±0.05	0.36±0.03

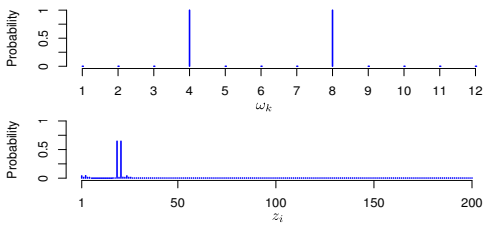

Figure 3: Avg. Posterior Prob. for $\omega_k = 1$ and $z_i = 1$.

The average results obtained are displayed in Table 1. The best method is DMFS. This model makes the hypothesis most compatible with the the data. The differences of DMFS with respect to the other methods are statistically significant (p-value $< 5\%$ using a paired Student's T test). MFS_{Dep} also performs well since the hypothesis made is also very flexible. DM and RMFS perform poorly in general. The reason for this is that DM is unable to model outlier tasks [6]. It can only model outlier features. Similarly, RMFS is unable to model outlier features [11]. Furthermore, in RMFS outlier tasks cannot be sparse. All the model coefficients of these tasks are expected to be relevant. Another reason for the bad behavior of DM and RMFS is that the norms employed by these methods cannot provide very sparse solutions without shrinking relevant coefficients [8, 16]. 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.

5 Conclusions

Most methods for multi-task feature selection assume jointly relevant and irrelevant features across tasks. This may be too restrictive in practice. In this work we have proposed a 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. Exact inference is infeasible under the proposed prior. However, a closed-form expectation propagation algorithm can be used for approximate inference. A model using such a prior has been evaluated showing gains in the prediction performance and in the identification of relevant features. Such a prior is also useful to better understand the data by identifying outlier tasks and outlier features.

Acknowledgement: D.H.L. is supported by MCyT and by CAM (projects TIN2010-21575-C02-02, TIN2013-42351-P and S2013/ICE-2845). J.M.H.L acknowledges support from the Rafael del Pino Foundation.

References

[1] J. E. Vogt and V. Roth. 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, pages 252–261, 2010.

[2] D. Hernández-Lobato, J. M. Hernández-Lobato, T. Helleputte, and P. Dupont. Expectation propagation for Bayesian multi-task feature selection. In *European Conference on Machine Learning*, volume 6321, pages 522–537, 2010.

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

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

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

[6] A. Jalali, P. Ravikumar, S. Sanghavi, and C. Ruan. A dirty model for multi-task learning. In *Advances in Neural Information Processing Systems 23*, pages 964–972. 2010.

[7] Thomas Minka. *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, MIT, 2001.

[8] C.M. Carvalho, N.G. Polson, and J.G. Scott. Handling sparsity via the horseshoe. *Journal of Machine Learning Research W&CP*, 5:73–80, 2009.

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

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

[11] P. Gong, J. Ye, and C. Zhang. Robust multi-task feature learning. In *18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 895–903. ACM, 2012.

[12] D. Hernández-Lobato and J. M. Hernández-Lobato. Learning feature selection dependencies in multi-task learning. In *Advances in Neural Information Processing Systems 26*, pages 746–754. 2013.

[13] Y. Xue, X. Liao, L. Carin, and B. Krishnapuram. Multi-task learning for classification with Dirichlet process priors. *Journal of Machine Learning Research*, 8:35–63, 2007.

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

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

[16] D. Hernández-Lobato, J. M. Hernández-Lobato, and P. Dupont. Generalized spike-and-slab priors for bayesian group feature selection using expectation propagation. *Journal of Machine Learning Research*, 14:1891–1945, 2013.