UAM UNIVERSIDAD AUTÓNOMA DE MADRID · HARVARD UNIVERSITY · UNIVERSITY OF CAMBRIDGE

A Probabilistic Model for Dirty Multi-task Feature Selection

Daniel Hernández-Lobato[1], José Miguel Hernández-Lobato[2,3], Zoubin Ghahramani[3]

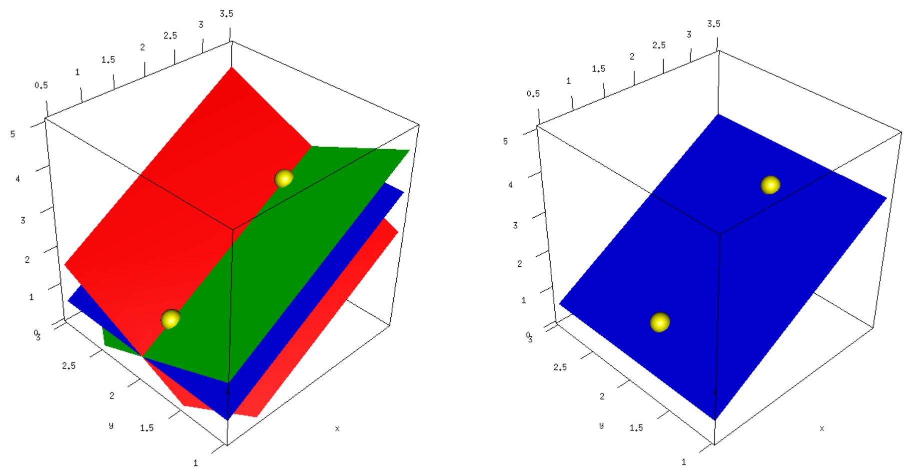
[1]Universidad Autónoma de Madrid. [2]Harvard University. [3]Cambridge University.

ICML Lille — International Conference on Machine Learning

1. Introduction

We focus on linear regression problems that are **under-determined**, *i.e.*, we have the same or more attributes than observations ($n \leq d$).

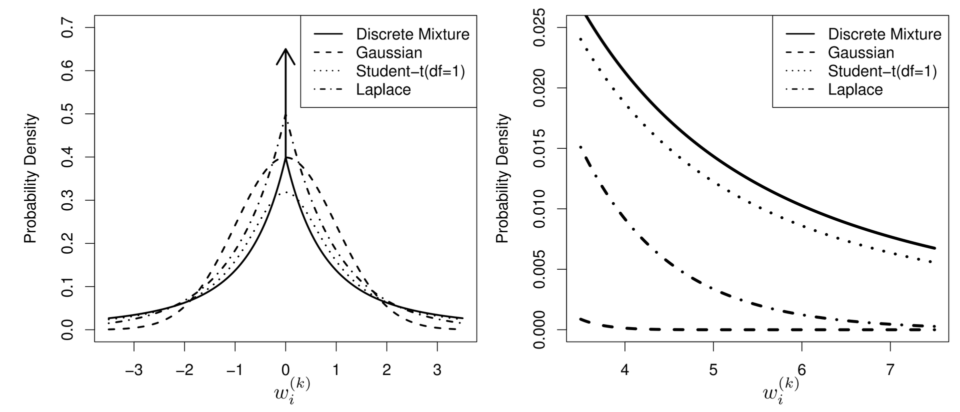
$$\mathbf{y} = \mathbf{X}\mathbf{w} + \epsilon\,, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}\sigma^2)\,.$$



A typical regularization assumes **sparsity** in $\mathbf{w}$.

2. Sparsity Assumption

Introduced by setting a **sparse enforcing prior** for $\mathbf{w}$, *e.g.*, Laplace, Student's T, Horseshoe or spike-and-slab.



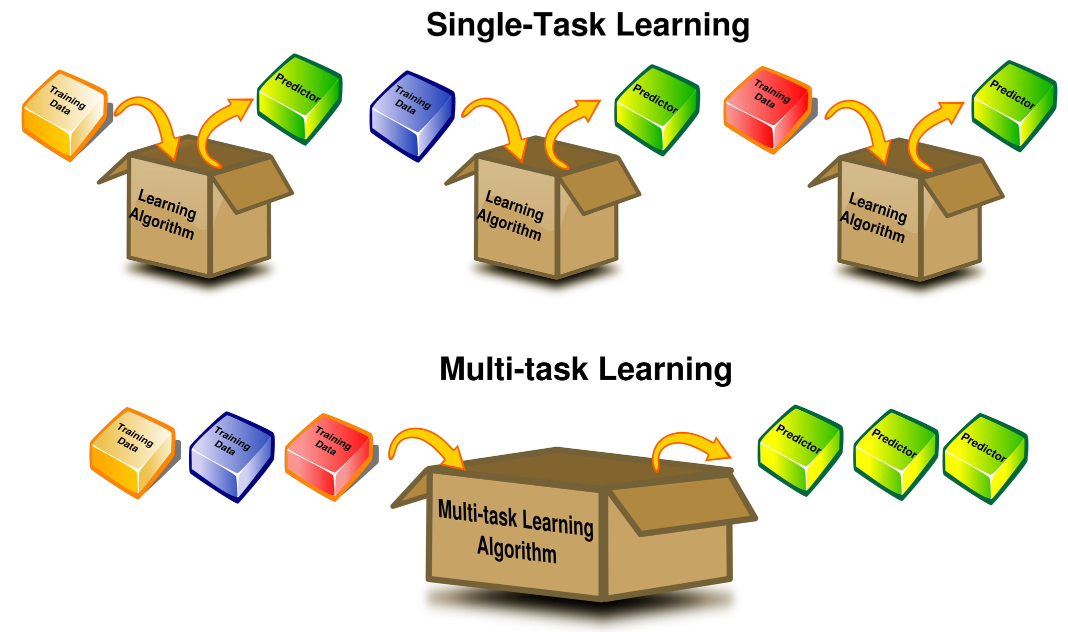
Discrete-mixture prior: $p(w_i) = \rho\delta_0 + (1-\rho)\pi(w_i)$.

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

is the Strawderman-Bergen prior which has a **closed form convolution** with the Gaussian distribution.

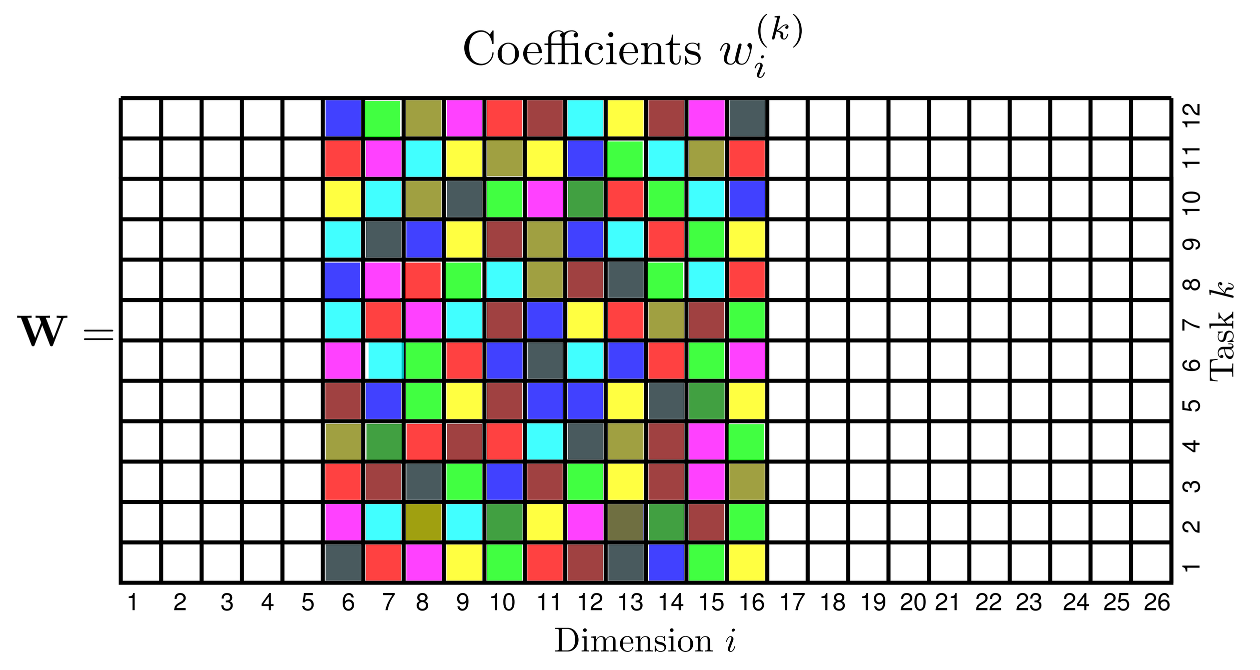
3. Multi-task Learning

There may be several learning tasks available for induction. Multi-task methods try to exploit **similarities** among tasks to improve the induction process.



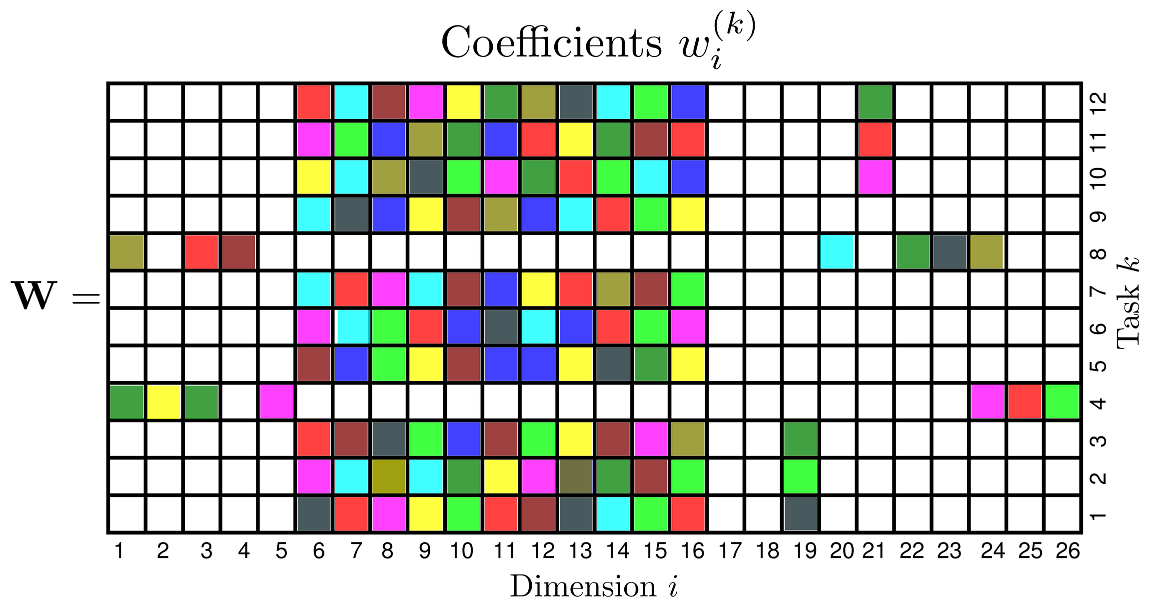
4. Typical Hypothesis

Tasks **share relevant and irrelevant** features.



5. Something More Reasonable

Outlier tasks and **outlier features**.



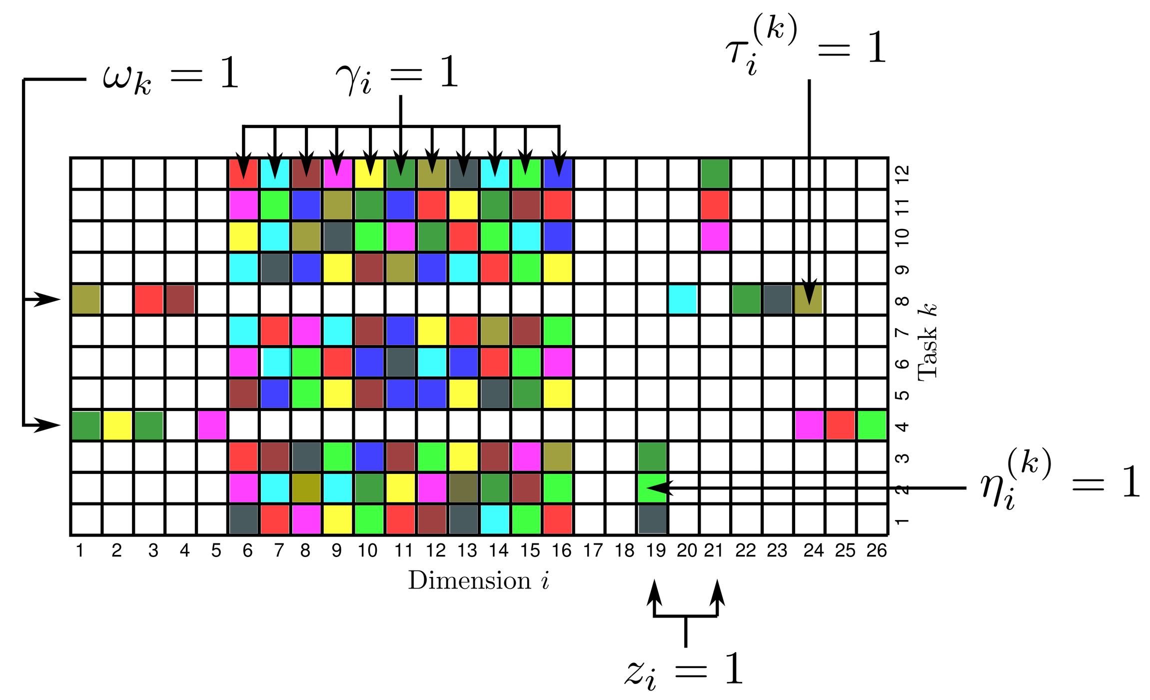
6. Robust Prior Distribution

Define $\Omega = \{\mathbf{z}, \omega, \gamma, \{\tau^{(k)}\}_{k=1}^K, \{\eta^{(k)}\}_{k=1}^K\}$. The prior for $\mathbf{W}$ is $p(\mathbf{W}|\Omega) = \prod_{i=1}^d \prod_{k=1}^K p(w_i^{(k)}|\Omega)$, with $p(w_i^{(k)}|\Omega) =$

$$\Bigg\{ \overbrace{\Big[\underbrace{\pi\left(w_i^{(k)}\right)^{\tau_i^{(k)}} \delta_0^{1-\tau_i^{(k)}}}_{\text{Outlier task}} \Big]^{\omega_k} \Big[\underbrace{\pi\left(w_i^{(k)}\right)^{\gamma_i} \delta_0^{1-\gamma_i}}_{\text{Not outlier task}} \Big]^{1-\omega_k}}^{\text{Not outlier feature}} \Bigg\}^{1-z_i} \times$$

$$\times \Bigg\{ \overbrace{\pi\left(w_i^{(k)}\right)^{\eta_i^{(k)}} \delta_0^{1-\eta_i^{(k)}}}^{\text{Outlier feature}} \Bigg\}^{z_i},$$

where δ_0 is a point of probability mass at the origin.



5. Dirty Multi-task Feature Selection

Gaussian Likelihood

$$\mathbf{y}^{(k)} \sim \mathcal{N}(\mathbf{X}^{(k)}\mathbf{w}^{(k)}, 0, \mathbf{I}\sigma^2_{(k)})\,, \quad \forall k\,.$$

$$w_i^{(k)} \sim \text{RobustPrior}(z_i, \omega_k, \gamma_i, \tau_i^{(k)}, \eta_i^{(k)})\,, \quad \forall i,k\,.$$

$$\sigma^2_{(k)} \sim \text{InvGam}(5,5)\,, \quad \forall k\,.$$ Noise hyper-prior

Independence assumption

$$z_i \sim \text{Bernoulli}(\rho_z)\,, \quad \forall i\,,$$
$$\omega_k \sim \text{Bernoulli}(\rho_\omega)\,, \quad \forall k\,,$$
$$\gamma_i \sim \text{Bernoulli}(\rho_\omega)\,, \quad \forall i\,,$$
$$\tau_i^{(k)} \sim \text{Bernoulli}(\rho_\tau)\,, \quad \forall i,k\,,$$
$$\eta_i^{(k)} \sim \text{Bernoulli}(\rho_\eta)\,, \quad \forall i,k\,,$$

Prob. hyper-prior

$$\rho_z \sim \text{Beta}(1,1)\,,$$
$$\rho_\omega \sim \text{Beta}(1,1)\,,$$
$$\rho_\gamma \sim \text{Beta}(1,1)\,,$$
$$\rho_\tau \sim \text{Beta}(1,1)\,,$$
$$\rho_\eta \sim \text{Beta}(1,1)\,,$$

7. Expectation Propagation

Approximates each factor in $p(\mathbf{Y}, \mathbf{W}, \Omega, \rho, \sigma^2|\mathcal{X})$ with an unormalized distribution inside an **exponential family** $\mathcal{F}$.

$\mathcal{F} \rightarrow$ Gaussians on $\mathbf{W}$, Bernoullis on Ω, I. Gammas on σ^2 and Betas on ρ.

Example of factors that need approximation:

$$\overbrace{\mathcal{N}(y_n^{(k)}|(\mathbf{x}_n^{(k)})^{\mathsf{T}}\mathbf{w}^{(k)}, \sigma^2_{(k)})}^{\text{Exact Factor } f_n^{(k)}} \xrightarrow{\text{EP Proj.}} \underbrace{\tilde{s}_n^{(k)}\text{InvGam}(\sigma^2_{(k)}|\tilde{a}_n^{(k)}, \tilde{b}_n^{(k)})\,\mathcal{N}((\mathbf{x}_n^{(k)})^{\mathsf{T}}\mathbf{w}^{(k)}|\tilde{m}_n^{(k)}, \tilde{v}_n^{(k)})}_{\text{Approx. Factor } \tilde{f}_n^{(k)}}$$

$$\overbrace{p(w_i^{(k)}|z_i, \omega_k, \gamma_i, \tau_i^{(k)}, \eta_i^{(k)})}^{\text{Exact Factor } g_i^{(k)}} \xrightarrow{\text{EP Proj.}} \underbrace{\tilde{s}_i^{(k)}\mathcal{N}(w_i^{(k)}|\tilde{m}_i^{(k)}, \tilde{v}_i^{(k)})\,\text{Bern}(z_i|\tilde{\rho}_z^{(i,k)})\cdots}_{\text{Approx. Factor } \tilde{g}_i^{(k)}}$$

$$\overbrace{\text{Bernoulli}(z_i|\rho_z)}^{\text{Exact Factor } h_i} \xrightarrow{\text{EP Proj.}} \underbrace{\tilde{s}_i\text{Beta}(\rho_z|\tilde{a}_i, \tilde{b}_i)\,\text{Bern}(z_i|\tilde{p}_i)}_{\text{Approx. Factor } \tilde{h}_i}$$

8. Experiments with Synthetic Data

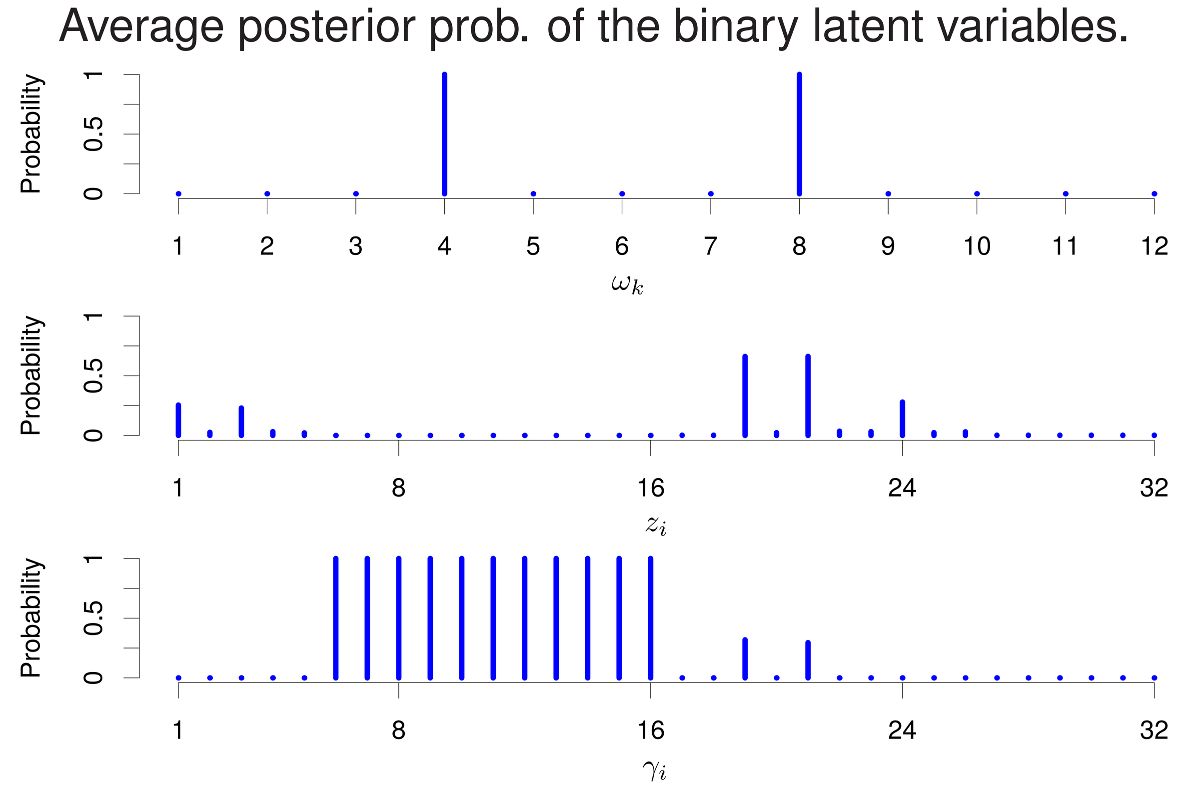
Several other experiments in the paper!

$N = 200$, $d = 2{,}000$ and we use for $\mathbf{W}$ the pattern above. $K = 12$, $\sigma^2_{(k)} = 0.5, \forall k$ and $w_i^{(k)} \sim$ Student(df = 5).

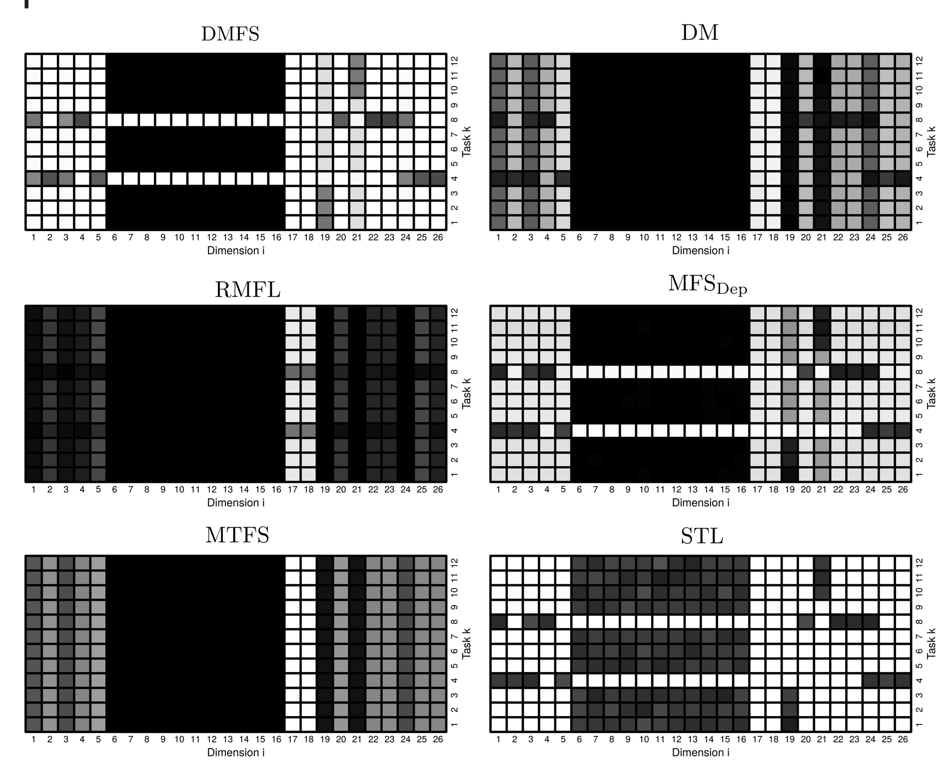
Method	Test RMSE	Rec. Error	Training Time
DMFS	**0.73** ±**0.04**	**0.22**±**0.02**	21.29±0.2
DM	0.86±0.05	0.50±0.03	150.35±10.0
RMFL	0.90±0.05	0.56±0.03	95.42±5.0
MFS_{Dep}	0.77±0.06	0.32±0.04	$\mathbf{2\cdot10^3 \pm 4\cdot10^2}$
MFS	0.81±0.06	0.37±0.04	6.7±1.7
STL	0.78±0.07	0.33±0.06	**4.76**±**0.4**

MFS and STL are **particular cases** of DMFS.

Average posterior prob. of the binary latent variables.



Average prob. that each coefficient is different from zero.



9. Summary and Conclusions of the Research Work

(1) - Most methods for multi-task feature selection assume jointly relevant and irrelevant features, which may be **too restrictive**. **(2)** - A robust prior allows tasks with **specific** relevant and irrelevant coefficients, and features to be **arbitrarily** relevant or irrelevant. **(3)** - Exact inference is infeasible under such a prior. However, a **quadrature-free** expectation propagation algorithm is possible. **(4)** - Several experiments show **gains** in the prediction performance and in the identification of relevant features for prediction. **(5)** - Our new prior is useful to **better understand** the data because it allows to identify outlier tasks and outlier features.

http://dhnzl.org/ daniel.hernandez@uam.es