Learning Feature Selection Dependencies in Multi-task Learning

Daniel Hernández-Lobato[1] and José Miguel Hernández-Lobato[2]

[1]Escuela Politécnica Superior, Universidad Autónoma de Madrid, Spain.
[2]Engineering Department, University of Cambridge, United Kingdom.

1. Introduction

- The strong regularization of **sparse linear regression** models is suitable in the large d but small n scenario to avoid over-fitting.
- The sparsity assumption can be introduced by carrying out Bayesian inference under a **sparsity enforcing prior** for the model coefficients.
- **Introducing dependencies** (groups) when determining relevant and irrelevant coefficients can improve the inference process.
- Most times these dependencies have to be **specified beforehand**.

2. Modelling Feature Selection Dependencies

Consider first a single task with some data in the form of n d-dimensional vectors $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\text{T}$ and targets $\mathbf{y} = (y_1, \ldots, y_n)^\text{T}$. Assume $\mathbf{y} = \mathbf{X}\mathbf{w} + \epsilon$, with $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, *i.e.*,

$$p(\mathbf{y}|\mathbf{X},\mathbf{w}) = \mathcal{N}(\mathbf{y}|\mathbf{X},\mathbf{w})\,.$$

The prior for $\mathbf{w}$ is the **horseshoe** sparse enforcing prior:

$$p(\mathbf{w}) = \prod_{j=1}^d p(w_j)\,, \qquad p(w_j|\tau) = \int \mathcal{N}(0, \lambda_j^2\tau)\mathcal{C}^+(\lambda_j)d\lambda_j\,,$$

where τ controls the **level of sparsity** and $\mathcal{C}^+(\cdot)$ a truncated Cauchy.

We consider the following **alternative representation**:

$$p(\mathbf{w},\mathbf{u},\mathbf{v}) = \left[\prod_{j=1}^d \mathcal{N}(w_j|0, u_j^2/v_j^2)\right] \mathcal{N}(\mathbf{u}|\mathbf{0},\rho^2\mathbf{C})\mathcal{N}(\mathbf{v}|\mathbf{0},\gamma^2\mathbf{C})$$

where ρ^2 and γ^2 control the level of sparsity and u_j and v_j are latent variables related to the **importance of feature j**:

- The larger u_j^2 the more relevant the corresponding feature.
- The smaller v_j^2 the more irrelevant the corresponding feature.

$\mathbf{C}$ is a correlation matrix that introduces **dependencies** in the feature selection process. If $\mathbf{C} = \mathbf{I}$ and $\mathbf{u}$ and $\mathbf{v}$ are marginalized, the **original horseshoe prior** is obtained.

The form of $\mathbf{C}$ is constrained to depend on $m \ll d$ parameters only:

$$\mathbf{C} = \Delta\mathbf{M}\Delta\,, \quad \mathbf{M} = \mathbf{D} + \mathbf{P}\mathbf{P}^\text{T}\,, \quad \Delta = \text{diag}(1/\sqrt{M_{11}}, \ldots, 1/\sqrt{M_{dd}})\,,$$

where $\mathbf{P}$ is an arbitrary matrix of size $d \times m$ which determines $\mathbf{C}$.

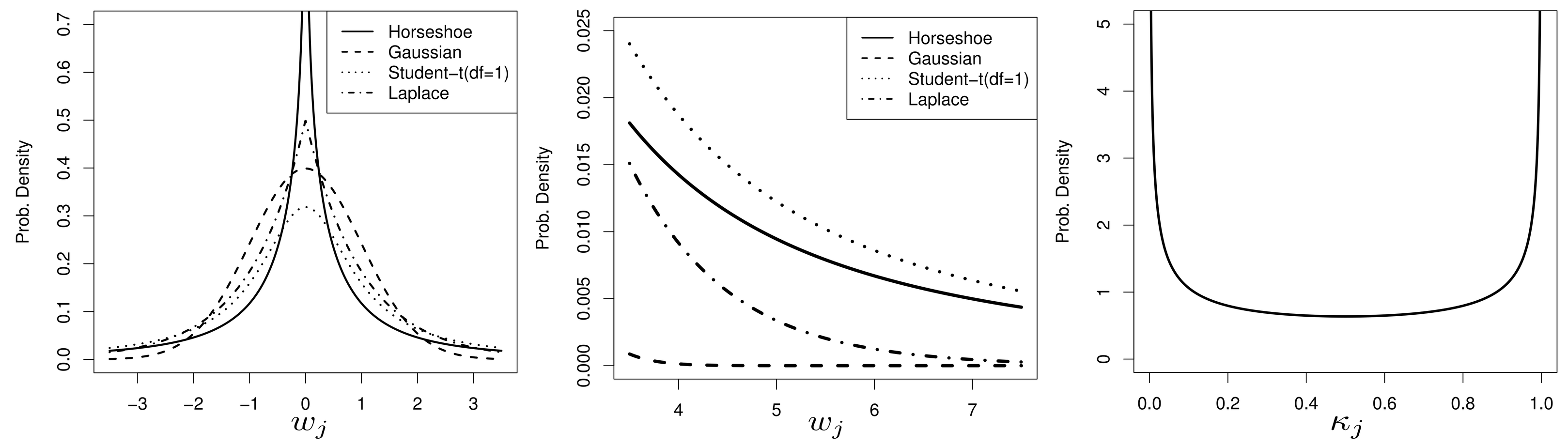
The joint **posterior** for the latent variables $\mathbf{z} = (\mathbf{w},\mathbf{u},\mathbf{v})$

$$p(\mathbf{z}|\mathbf{y},\mathbf{X}) = \frac{p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w},\mathbf{u},\mathbf{v})}{p(\mathbf{y})}\,,$$

is **intractable** and we have to resort to approximate inference.

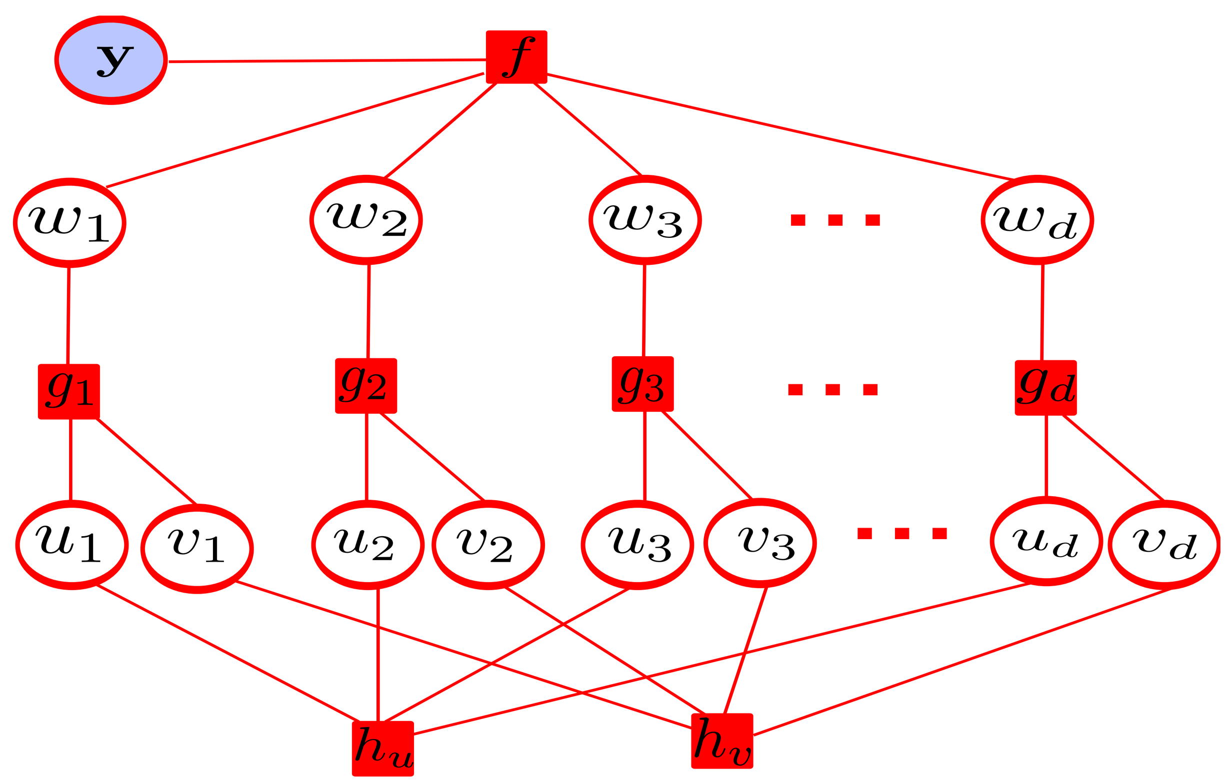
3. Horseshoe Prior Distribution

Assume $\tau = \sigma^2 = 1$ and $\mathbf{X} = \mathbf{I}$. Define $\kappa_j = 1/(1+\lambda_j^2)$. Then, the posterior mean of w_j is $(1-\kappa_j)y_j$, where κ_j is a shrinkage coefficient. If $\kappa_j = 0$ ($\kappa_j = 1$) there is **no shrinkage** (**total shrinkage**).



The prior distribution for κ_j tends to infinity both at **0** and **1**.

4. Factor Graph of the Probabilistic Model



$f(\cdot)$ corresponds to $\mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w},\sigma^2\mathbf{I})$, each $g_j(\cdot)$ to $\mathcal{N}(w_j|0,u_j^2/v_j^2)$, $h_u(\cdot)$ and $h_v(\cdot)$ to $\mathcal{N}(\mathbf{u}|\mathbf{0},\rho^2\mathbf{C})$ and $\mathcal{N}(\mathbf{v}|\mathbf{0},\gamma^2\mathbf{C})$, respectively.

5. Approximate Inference and Multi-task Extension

- Approximate inference is implemented by **expectation propagation**.
- Approximates each non-Gaussian factor g_j by a Gaussian factor $\tilde{g}_j$.
- **EP** provides an estimate of the marginal likelihood $p(\mathbf{y}|\sigma^2,\rho^2,\gamma^2,\mathbf{C})$.
- This estimate is **maximized** with respect to $\mathbf{C}$ using gradient ascent.

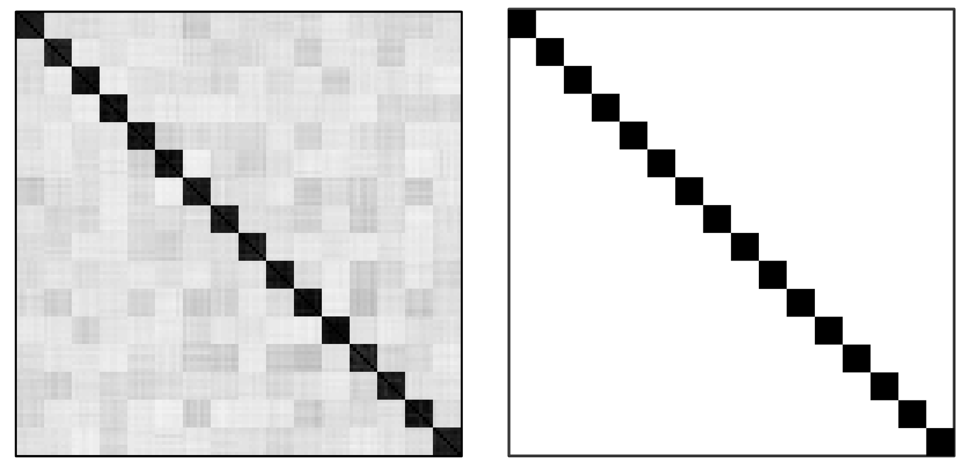
The EP updates cannot be computed in closed form but can be evaluated using **one-dimensional quadrature**. The cost is $\mathcal{O}(n^2d)$.

- A multi-task extension is obtained by sharing $\mathbf{C}$ across learning tasks.
- The K tasks may have **different relevant attributes or coefficients**.
- The **EP** estimate of $\prod_{k=1}^K p(\mathbf{y}_k|\sigma_k^2,\rho_k^2,\gamma_k^2,\mathbf{C})$ is maximized to find $\mathbf{C}$.

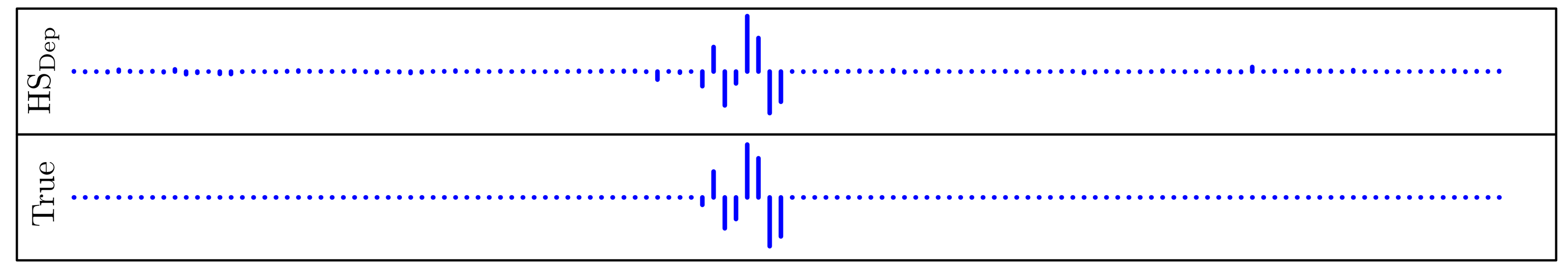
6. Reconstruction of Sparse Signals

$K = 64$ signals are generated using a particular **sparsity pattern**. They are then reconstructed from a set of **Gaussian measurements**.

Method	Error
HS_{ST}	0.29±0.01
HS_{MT}	0.38±0.03
SS_{MT}	0.77±0.01
DM	0.37±0.01
BM	0.24±0.02
HS_{Dep}	0.21±0.01



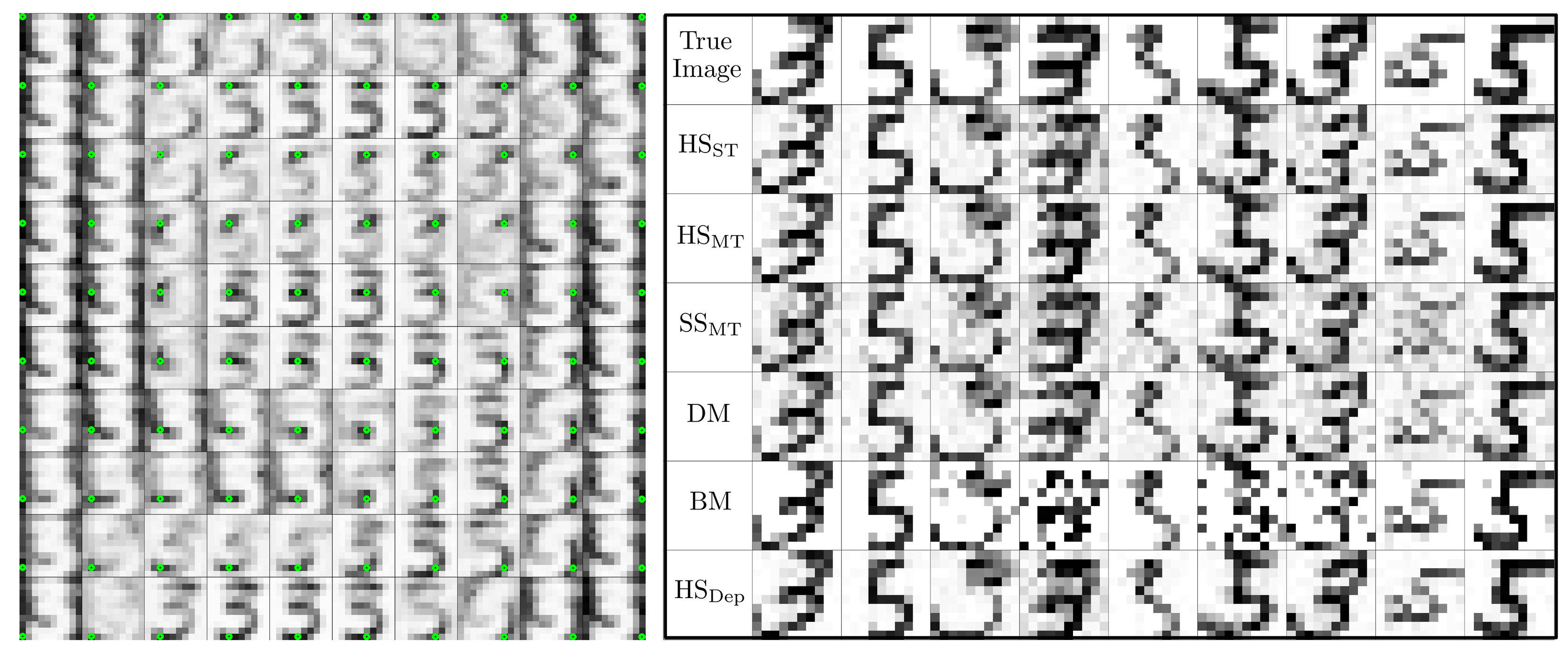
(left) Reconstruction Errors. (middle) Avg. Estimated Correlations. (right) True Correlations.



7. Reconstruction of Images of Hand-written Digits

$K = 100$ images corresponding to the digits **5 and 3** from the MNIST dataset are reconstructed from a set of **Gaussian measurements**.

	HS_{ST}	HS_{MT}	SS_{MT}	DM	BM	HS_{Dep}
Error	0.36±0.02	0.25±0.02	0.39±0.01	0.37±0.01	0.52±0.03	0.20±0.01



8. Conclusions

- **Dependencies** in the feature selection process can **improve** the induction process of the model coefficients in sparse linear models.
- Dependencies can be **learnt from the training data** in a multi-task setting where the tasks share a **common dependency structure**.
- The horseshoe prior can be **easily adapted** for this purpose and **approximate inference** can be efficiently carried out using **EP**.