A Probabilistic Model for Dirty Multi-task Feature Selection

Daniel Hernández–Lobato[1],

May 7, 2015

joint work with

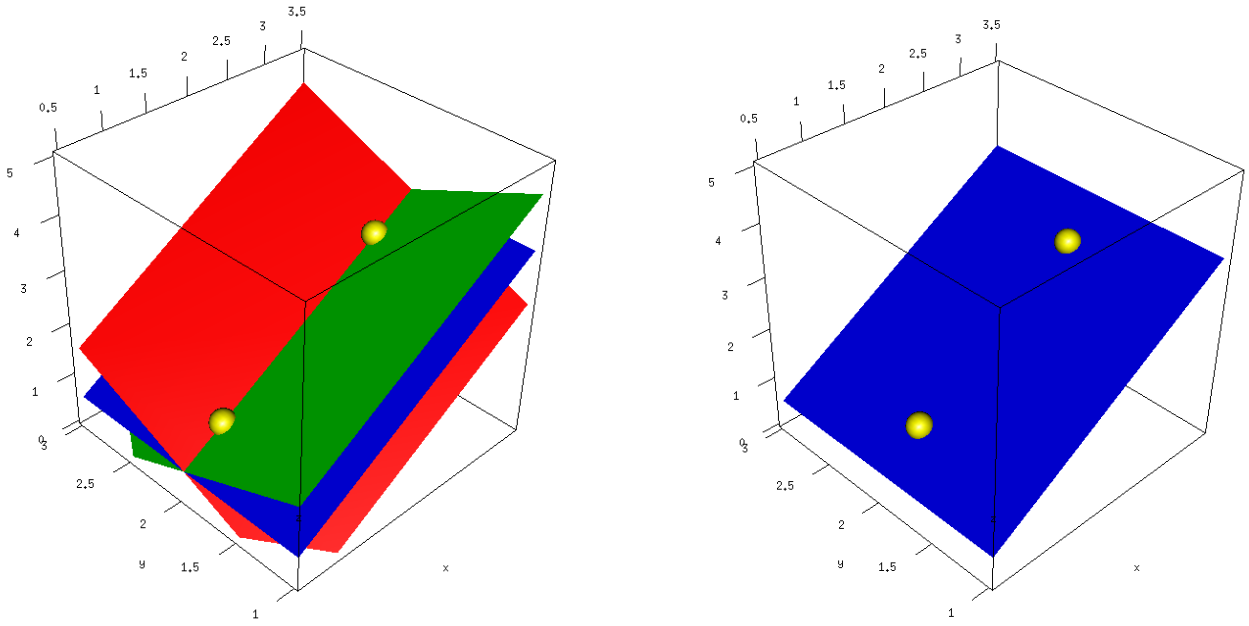
José Miguel Hernández-Lobato[2] and Zoubin Ghahramani[3]

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

Introduction

Linear regression problems are **under-determined** when we have the same or more attributes than observations ($n \leq d$).

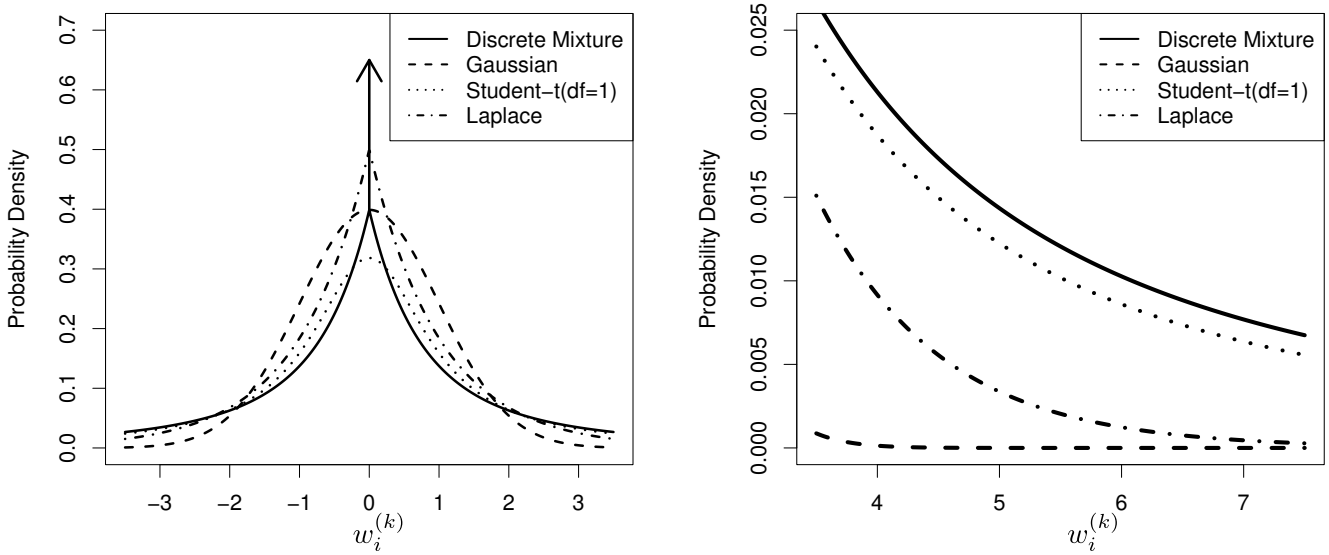
$$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}\,, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}\sigma^2)\,.$$



A typical regularization assumes **sparsity** in $\mathbf{w}$, i.e., most coefficients are equal to zero (Johnstone & Titterington, 2009).

Induction under the sparsity assumption

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



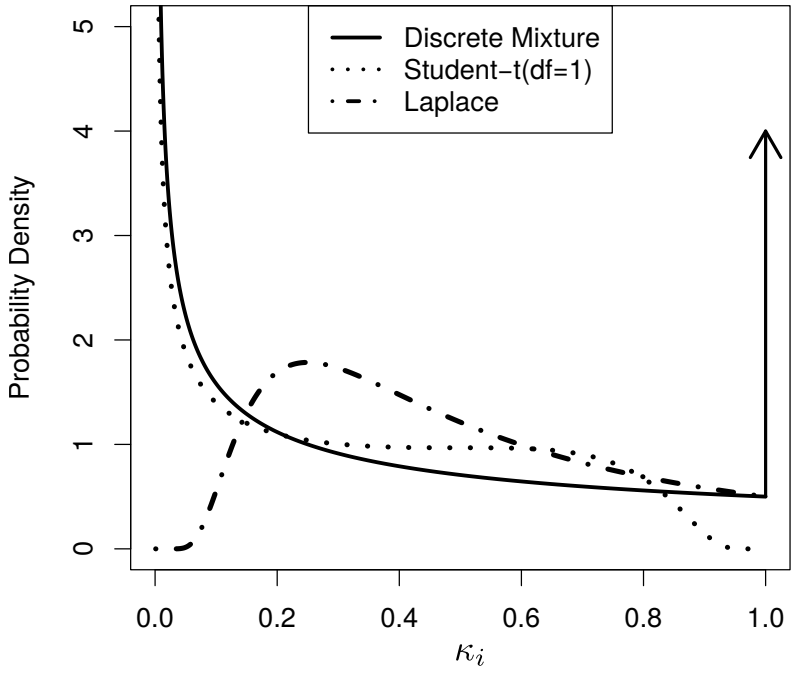
Discrete-mixture prior: $p(w_i) = \rho\delta_0 + (1-\rho)\pi(w_i)$.

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

is the Strawderman-Berger prior which has a **closed form convolution** with the Gaussian (Strawderman, 1971; Berger, 1980).

Shrinkage analysis

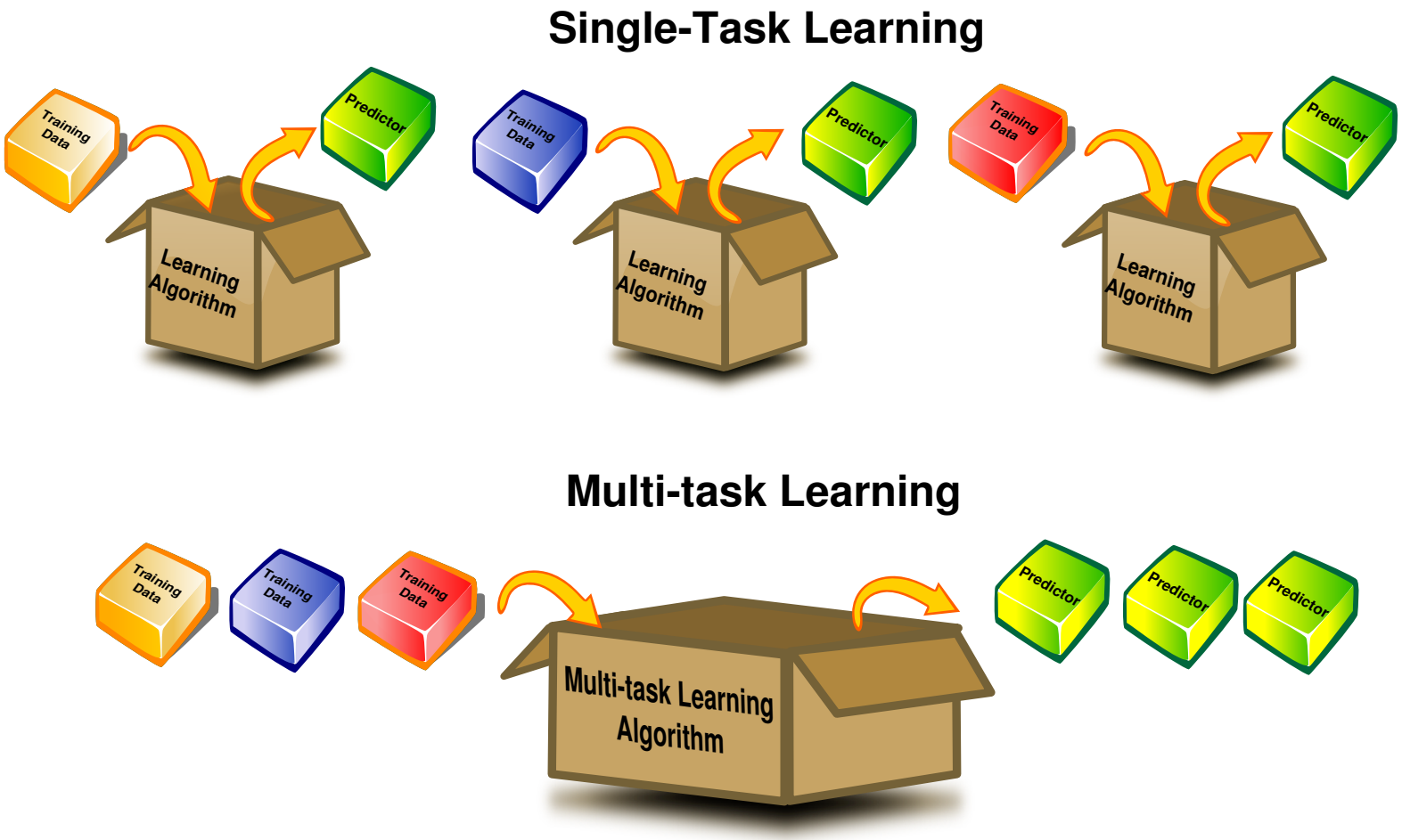
Assume that $\sigma^2 = 1$, $\mathbf{X} = \mathbf{I}$ and define $\kappa_j = 1/(1+\lambda_j^2)$. Then the posterior mean for w_j is $(1-\kappa_j)y_j$, where κ_j is a random **shrinkage coefficient**.



The discrete mixture is very convenient for sparse induction and can be considered as a **gold standard** (Carvalho et al., 2009).

Multi-task learning

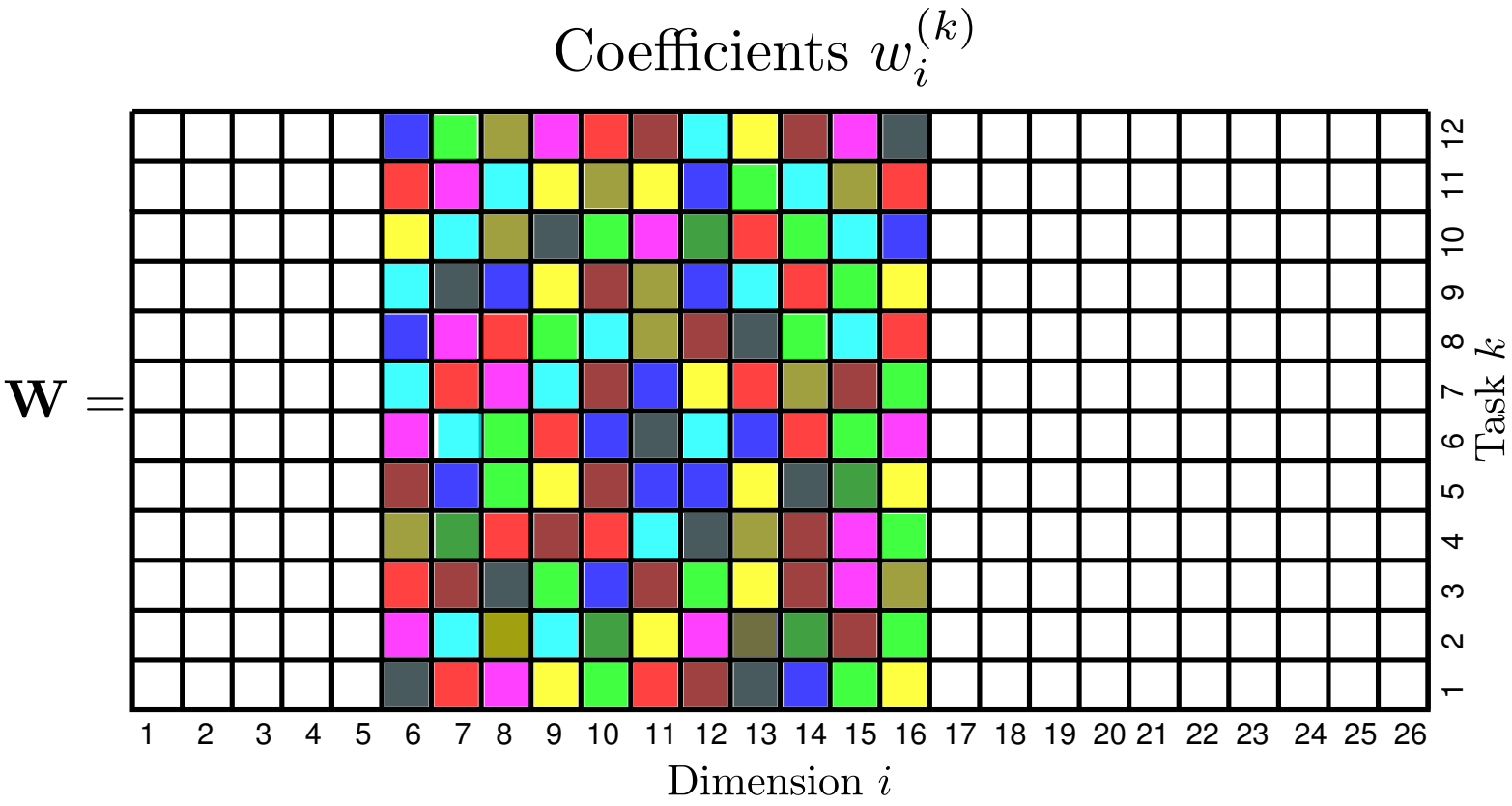
There may be several learning tasks available for induction.



Multi-task methods try to exploit **similarities** among tasks.

Typical hypothesis under the sparsity assumption

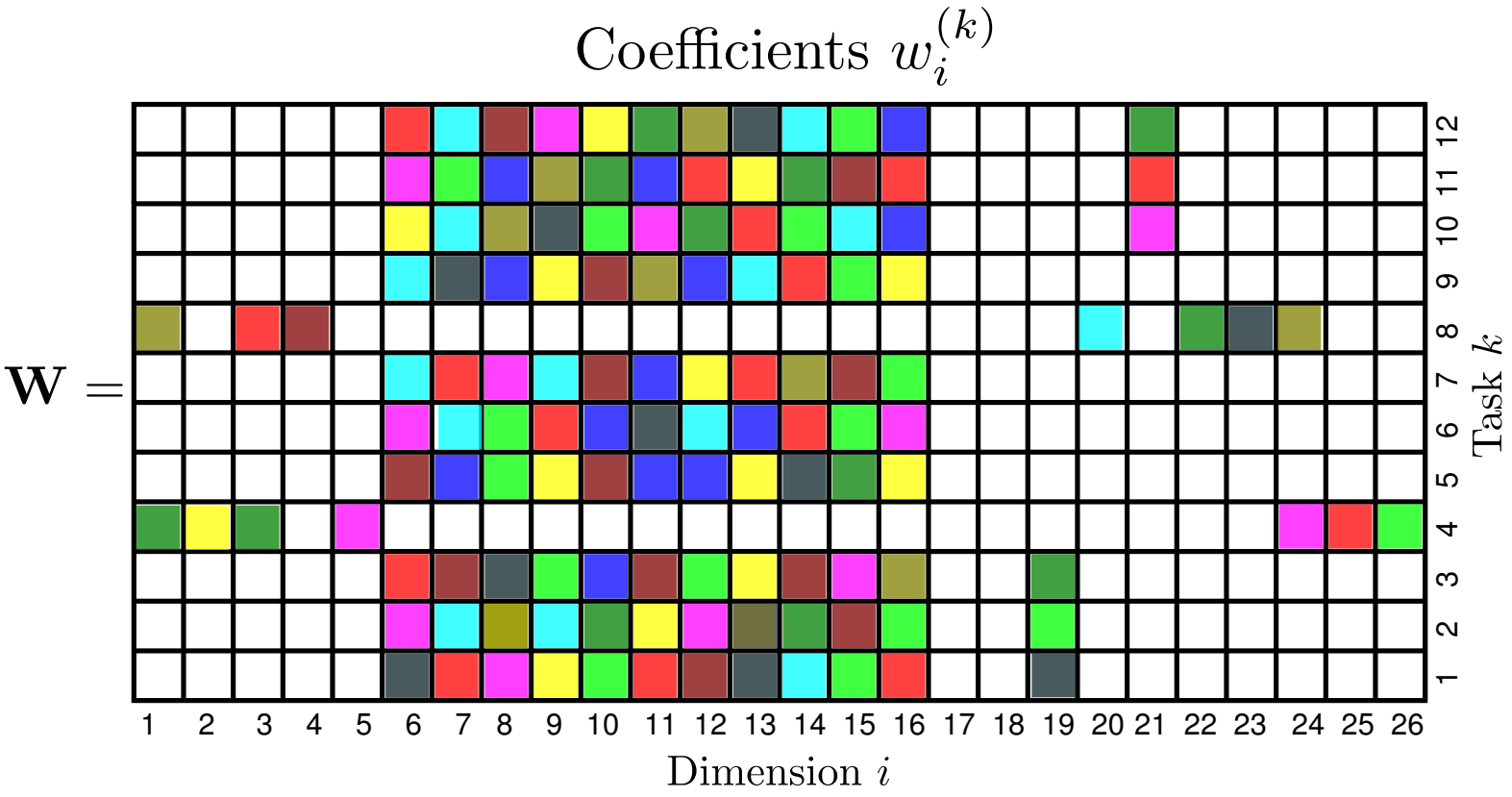
Tasks **share relevant and irrelevant** features.



This is the assumption made in, *e.g.*, (Hernández-Lobato et al., 2010; Jebara, 2004; Obozinski et al., 2009; Vogt & Roth, 2010; Xiong et al., 2007; Argyriou et al., 2007).

A more reasonable scenario

Most tasks **share relevant and irrelevant** features, but there are a few **outlier tasks** and a few **outlier features**.



How can we account for all this?

Back to Experiments

Robust prior distribution for $\mathbf{W}$ (I)

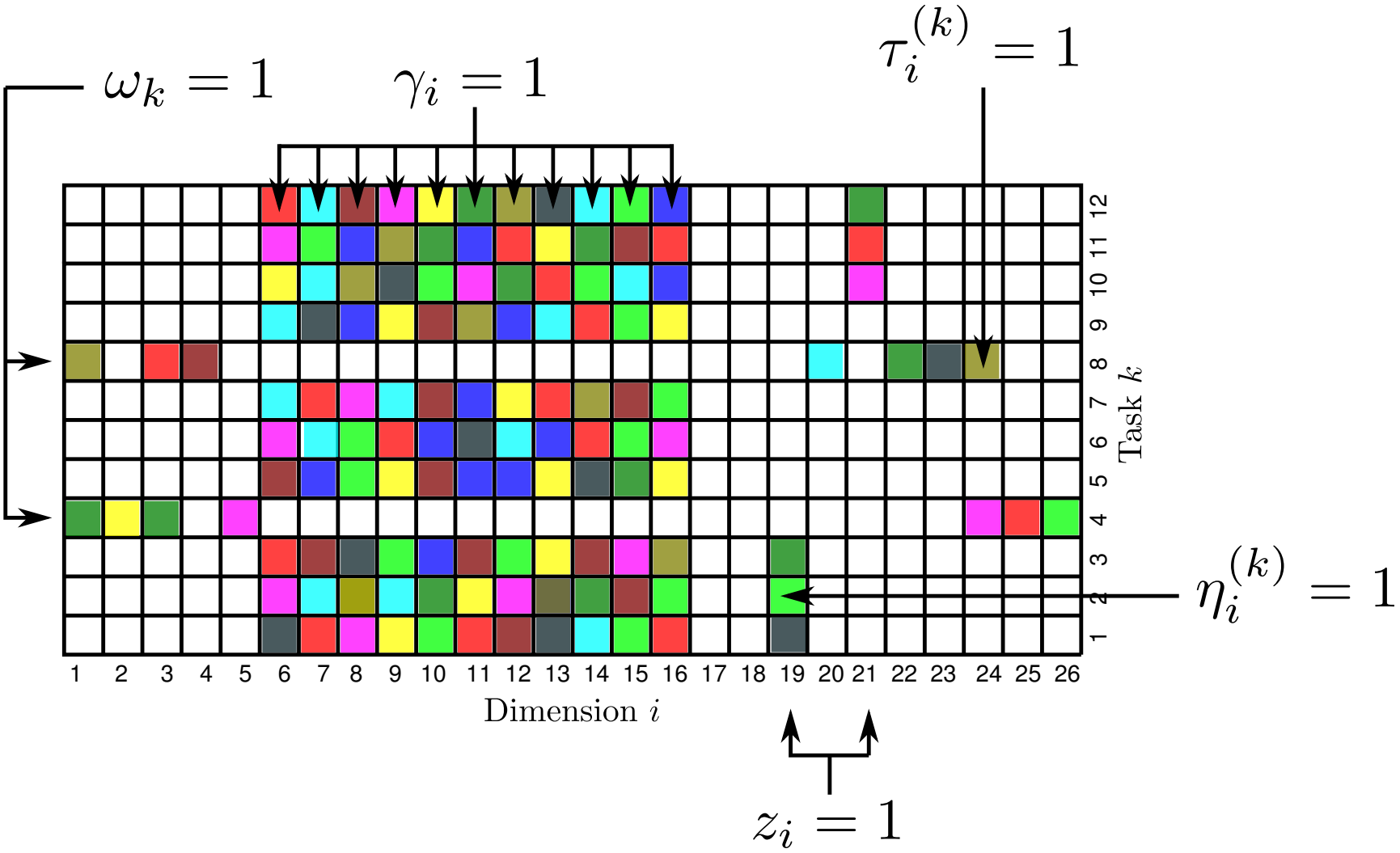
We introduce the following set of binary latent variables:

- 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 ($\eta_i^{(k)} = 0$).

We **summarize** all these variables in $\boldsymbol{\Omega} = \{\mathbf{z}, \boldsymbol{\omega}, \boldsymbol{\gamma}, \{\boldsymbol{\tau}^{(k)}\}_{k=1}^K, \{\boldsymbol{\eta}^{(k)}\}_{k=1}^K\}$.

Robust prior distribution for $\mathbf{W}$ (II)

Relation of binary latent variables for the example:



Robust prior distribution for $\mathbf{W}$ (III)

Given $\mathbf{\Omega}$ the prior for $\mathbf{W}$ is:

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

where $p(w_i^{(k)}|\mathbf{\Omega}) =$

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

Under this prior $w_i^{(k)}$ is different from zero iff:

1. It corresponds to an outlier feature ($z_i = 1$) relevant for task k ($\eta_i^{(k)} = 1$).
2. 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 ($\tau_i^{(k)} = 1$).
3. 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 ($\gamma_i = 1$).

Dirty multi-task feature selection model (DMFS)

Gaussian Likelihood

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

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

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

Hyper-prior for the noise

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

Independence among variables

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

Hyper-prior for each prob.

Task k and dimension i.

Inference, prediction and relevant features

Define $\boldsymbol{\mathcal{X}} = \{\mathbf{X}^{(k)}\}_{k=1}^{K}$ and $\boldsymbol{\rho} = (\rho_z, \rho_\omega, \rho_\gamma, \rho_\tau, \rho_\eta)^{\mathrm{T}}$.

The **posterior** is:

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

The **predictive distribution** for y_{new} given $\mathbf{x}_{\text{new}}$ of task k is:

$$p(y_{\text{new}} | \mathbf{x}_{\text{new}}) = \sum_{\boldsymbol{\Omega}} \int \mathcal{N}(y_{\text{new}} | \mathbf{x}_{\text{new}}^{\mathrm{T}} \mathbf{w}^{(k)}, \sigma_{(k)}^2) p(\mathbf{W}, \boldsymbol{\Omega}, \boldsymbol{\rho}, \boldsymbol{\sigma}^2 | \mathbf{Y}, \boldsymbol{\mathcal{X}}) d\mathbf{W} d\boldsymbol{\rho} d\boldsymbol{\sigma}^2 .$$

The probability that $w_i^{(k)}$ is **different from zero** is:

$$p(w_i^{(k)} \neq 0 | \mathbf{Y}, \boldsymbol{\mathcal{X}}) = p(\{z_i = 1 \cap \eta_i^{(k)} = 1\} \cup \{z_i = 0 \cap \omega_k = 1 \cap \tau_i^{(k)} = 1\} \cup \\ \cup \{z_i = 0 \cap \omega_k = 0 \cap \gamma_i = 1\} | \mathbf{Y}, \boldsymbol{\mathcal{X}}).$$

All these computations are intractable in practice!

Inference, prediction and relevant features

Define $\boldsymbol{\mathcal{X}} = \{\mathbf{X}^{(k)}\}_{k=1}^{K}$ and $\boldsymbol{\rho} = (\rho_z, \rho_\omega, \rho_\gamma, \rho_\tau, \rho_\eta)^{\mathrm{T}}$.

The **posterior** is:

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

The **predictive distribution** for y_{new} given $\mathbf{x}_{\text{new}}$ of task k is:

$$p(y_{\text{new}} | \mathbf{x}_{\text{new}}) = \sum_{\boldsymbol{\Omega}} \int \mathcal{N}(y_{\text{new}} | \mathbf{x}_{\text{new}}^{\mathrm{T}} \mathbf{w}^{(k)}, \sigma_{(k)}^2) p(\mathbf{W}, \boldsymbol{\Omega}, \boldsymbol{\rho}, \boldsymbol{\sigma}^2 | \mathbf{Y}, \boldsymbol{\mathcal{X}}) d\mathbf{W} d\boldsymbol{\rho} d\boldsymbol{\sigma}^2 .$$

The probability that $w_i^{(k)}$ is **different from zero** is:

$$p(w_i^{(k)} \neq 0 | \mathbf{Y}, \boldsymbol{\mathcal{X}}) = p(\{z_i = 1 \cap \eta_i^{(k)} = 1\} \cup \{z_i = 0 \cap \omega_k = 1 \cap \tau_i^{(k)} = 1\} \cup \\ \cup \{z_i = 0 \cap \omega_k = 0 \cap \gamma_i = 1\} | \mathbf{Y}, \boldsymbol{\mathcal{X}}).$$

All these computations are intractable in practice!

Expectation propagation (I)

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

$\mathcal{F} \rightarrow$ Gaussians on $\mathbf{W}$, Bernoullis on $\boldsymbol{\Omega}$, I. Gammas on $\boldsymbol{\sigma}^2$ and Betas on $\boldsymbol{\rho}$.

Example of factors that need approximation:

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

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

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

Factors such as $\mathrm{InvGam}(\sigma_{(k)}^2 | 5, 5)$ and $\mathrm{Beta}(\rho_z | 1, 1)$ need not be approximated.

Expectation propagation (II)

- The approximate posterior $\mathcal{Q}$ is obtained by **normalizing** the joint where exact factors are replaced by their approximation.
- EP adjusts each $\tilde{g}_i^{(k)}$ by minimizing $\mathrm{KL}[g_i^{(k)}\mathcal{Q}^{\setminus(i,k)}||\tilde{g}_i^{(k)}\mathcal{Q}^{\setminus(i,k)}]$, where $\mathcal{Q}^{\setminus(i,k)}$ is the distribution obtained from the ratio $\mathcal{Q}/\tilde{g}_i^{(k)}$.
- The minimization of $\mathrm{KL}[g_i^{(k)}\mathcal{Q}^{\setminus(i,k)}||\tilde{g}_i^{(k)}\mathcal{Q}^{\setminus(i,k)}]$ is done by **matching moments** between the two prob. distributions.
- Let $Z_{i,k}$ be the normalization constant of $g_i^{(k)}\mathcal{Q}^{\setminus(i,k)}$. All moments can be obtained from the **derivatives** of $\log Z_{i,k}$.
- $\log Z_{i,k}$ can be evaluated analytically because $\pi(w_i^{(k)})$ has a **closed form convolution** with the Gaussian distribution.

Expectation propagation (II)

- The approximate posterior $\mathcal{Q}$ is obtained by **normalizing** the joint where exact factors are replaced by their approximation.
- EP adjusts each $\tilde{g}_i^{(k)}$ by minimizing $\mathrm{KL}[g_i^{(k)}\mathcal{Q}^{\setminus(i,k)}||\tilde{g}_i^{(k)}\mathcal{Q}^{\setminus(i,k)}]$, where $\mathcal{Q}^{\setminus(i,k)}$ is the distribution obtained from the ratio $\mathcal{Q}/\tilde{g}_i^{(k)}$.
- The minimization of $\mathrm{KL}[g_i^{(k)}\mathcal{Q}^{\setminus(i,k)}||\tilde{g}_i^{(k)}\mathcal{Q}^{\setminus(i,k)}]$ is done by **matching moments** between the two prob. distributions.
- Let $Z_{i,k}$ be the normalization constant of $g_i^{(k)}\mathcal{Q}^{\setminus(i,k)}$. All moments can be obtained from the **derivatives** of $\log Z_{i,k}$.
- $\log Z_{i,k}$ can be evaluated analytically because $\pi(w_i^{(k)})$ has a **closed form convolution** with the Gaussian distribution.

Expectation propagation (II)

- The approximate posterior $\mathcal{Q}$ is obtained by **normalizing** the joint where exact factors are replaced by their approximation.
- EP adjusts each $\tilde{g}_i^{(k)}$ by minimizing $\mathrm{KL}[g_i^{(k)}\mathcal{Q}^{\setminus(i,k)}||\tilde{g}_i^{(k)}\mathcal{Q}^{\setminus(i,k)}]$, where $\mathcal{Q}^{\setminus(i,k)}$ is the distribution obtained from the ratio $\mathcal{Q}/\tilde{g}_i^{(k)}$.
- The minimization of $\mathrm{KL}[g_i^{(k)}\mathcal{Q}^{\setminus(i,k)}||\tilde{g}_i^{(k)}\mathcal{Q}^{\setminus(i,k)}]$ is done by **matching moments** between the two prob. distributions.
- Let $Z_{i,k}$ be the normalization constant of $g_i^{(k)}\mathcal{Q}^{\setminus(i,k)}$. All moments can be obtained from the **derivatives** of $\log Z_{i,k}$.
- $\log Z_{i,k}$ can be evaluated analytically because $\pi(w_i^{(k)})$ has a **closed form convolution** with the Gaussian distribution.

Expectation propagation (II)

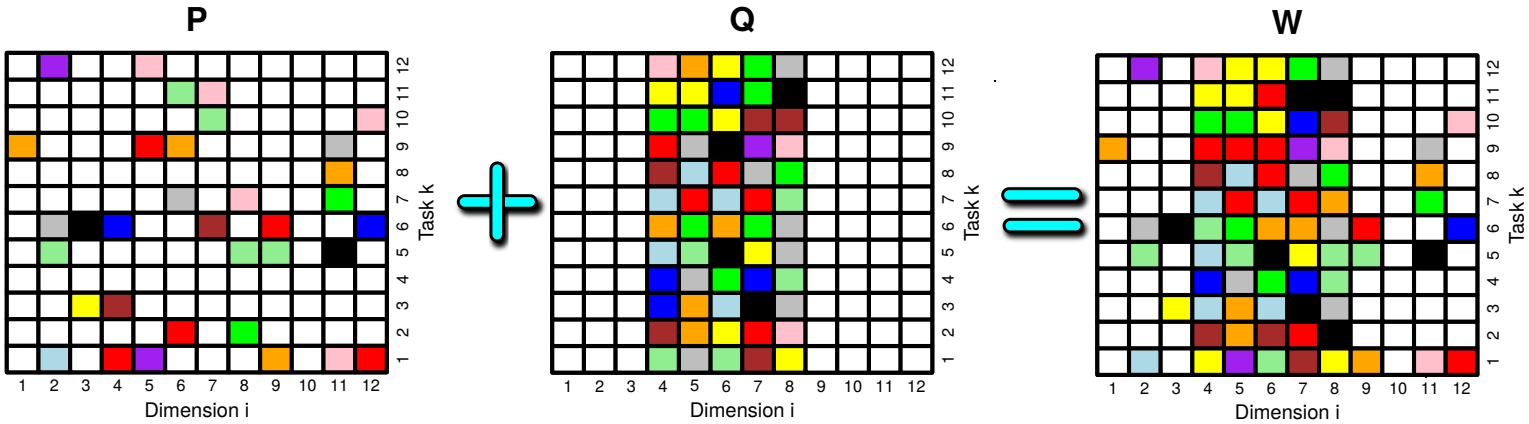
- The approximate posterior $\mathcal{Q}$ is obtained by **normalizing** the joint where exact factors are replaced by their approximation.
- EP adjusts each $\tilde{g}_i^{(k)}$ by minimizing $\mathrm{KL}[g_i^{(k)}\mathcal{Q}^{\setminus(i,k)}||\tilde{g}_i^{(k)}\mathcal{Q}^{\setminus(i,k)}]$, where $\mathcal{Q}^{\setminus(i,k)}$ is the distribution obtained from the ratio $\mathcal{Q}/\tilde{g}_i^{(k)}$.
- The minimization of $\mathrm{KL}[g_i^{(k)}\mathcal{Q}^{\setminus(i,k)}||\tilde{g}_i^{(k)}\mathcal{Q}^{\setminus(i,k)}]$ is done by **matching moments** between the two prob. distributions.
- Let $Z_{i,k}$ be the normalization constant of $g_i^{(k)}\mathcal{Q}^{\setminus(i,k)}$. All moments can be obtained from the **derivatives** of $\log Z_{i,k}$.
- $\log Z_{i,k}$ can be evaluated analytically because $\pi(w_i^{(k)})$ has a **closed form convolution** with the Gaussian distribution.

Expectation propagation (II)

- The approximate posterior $\mathcal{Q}$ is obtained by **normalizing** the joint where exact factors are replaced by their approximation.
- EP adjusts each $\tilde{g}_i^{(k)}$ by minimizing $\mathrm{KL}[g_i^{(k)}\mathcal{Q}^{\setminus(i,k)}||\tilde{g}_i^{(k)}\mathcal{Q}^{\setminus(i,k)}]$, where $\mathcal{Q}^{\setminus(i,k)}$ is the distribution obtained from the ratio $\mathcal{Q}/\tilde{g}_i^{(k)}$.
- The minimization of $\mathrm{KL}[g_i^{(k)}\mathcal{Q}^{\setminus(i,k)}||\tilde{g}_i^{(k)}\mathcal{Q}^{\setminus(i,k)}]$ is done by **matching moments** between the two prob. distributions.
- Let $Z_{i,k}$ be the normalization constant of $g_i^{(k)}\mathcal{Q}^{\setminus(i,k)}$. All moments can be obtained from the **derivatives** of $\log Z_{i,k}$.
- $\log Z_{i,k}$ can be evaluated analytically because $\pi(w_i^{(k)})$ has a **closed form convolution** with the Gaussian distribution.

Related methods (I) (DM)

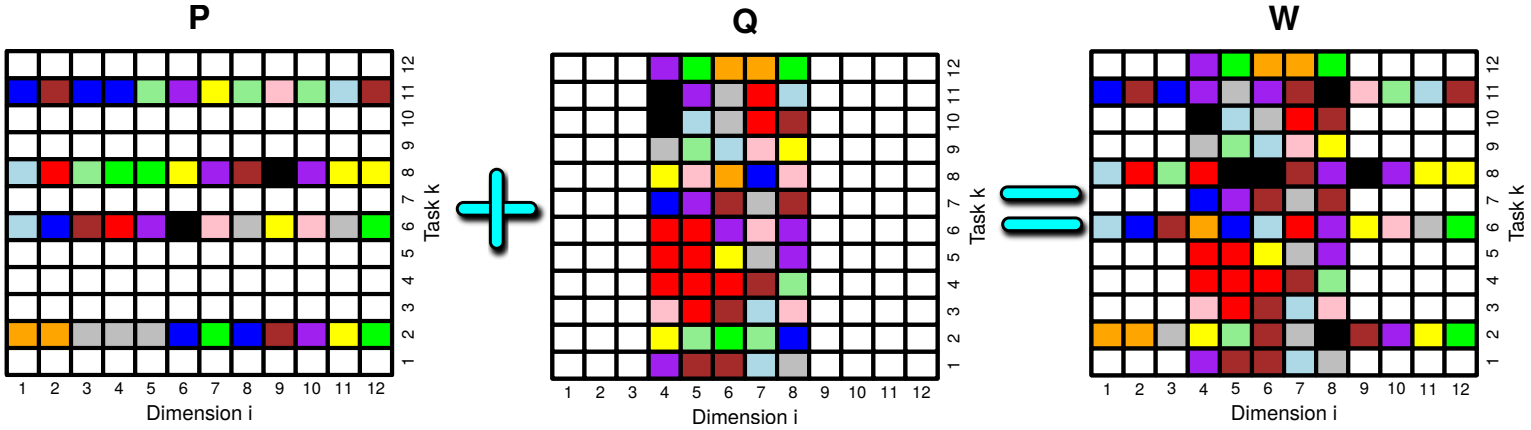
Dirty model: Jalali et al. (2010) assume $\mathbf{W} = \mathbf{P} + \mathbf{Q}$. They penalize $\mathbf{P}$ with the ℓ_1 norm and $\mathbf{Q}$ with the $\ell_{1,2}$ norm.



Assumes a few features jointly relevant **for all tasks** and some features that may be **specifically relevant** only for some tasks.

Related methods (II) (RMFL)

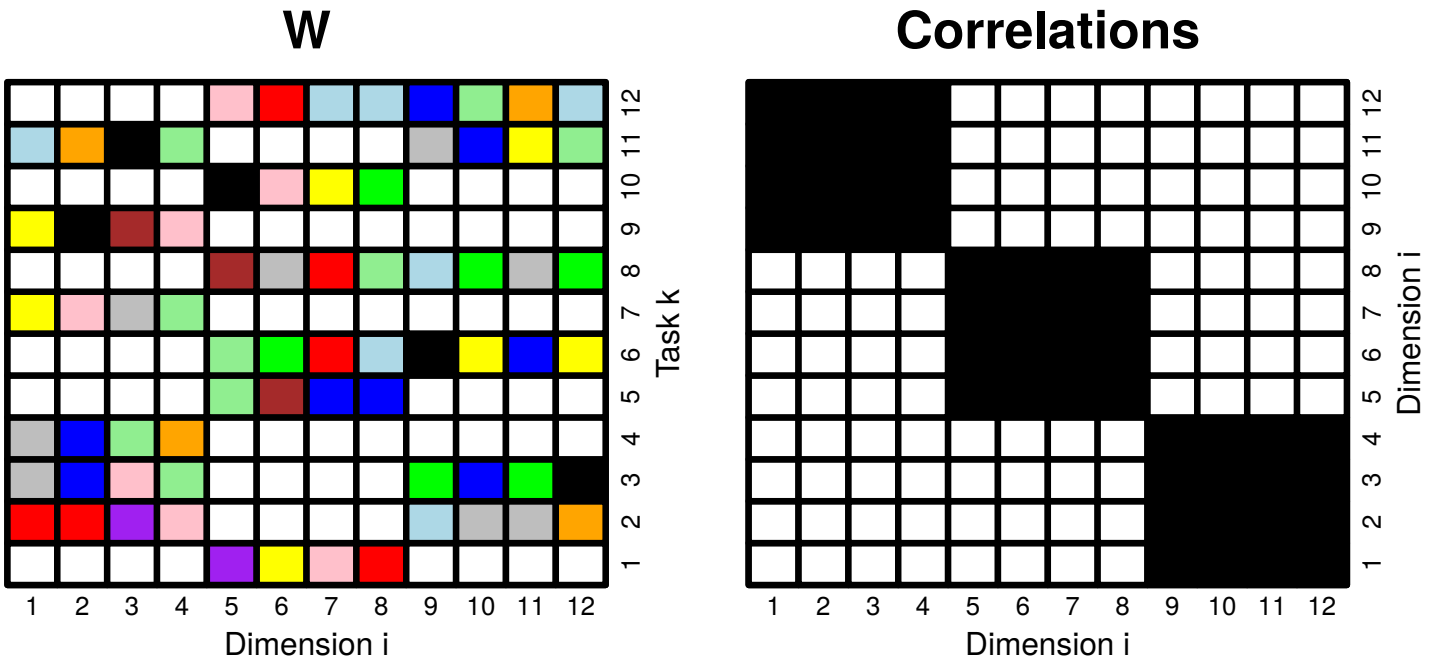
Robust multi-task feature learning: Gong et al. (2012) assume $\mathbf{W} = \mathbf{P} + \mathbf{Q}$. They penalize both $\mathbf{P}^{\mathrm{T}}$ and $\mathbf{Q}$ with the $\ell_{1,2}$ norm.



Assumes a few features jointly relevant **for all tasks** and **some outlier tasks** with all features relevant for prediction.

Related methods (III) ($\mathrm{MFS}_{\mathrm{Dep}}$)

Multi-task feature selection with dependencies: Common correlations are shared in the feature selection process of each task (Hernández-Lobato & Hernández-Lobato, 2013).



The relevant features for each task **may be different**.

Experiments with synthetic data (I)

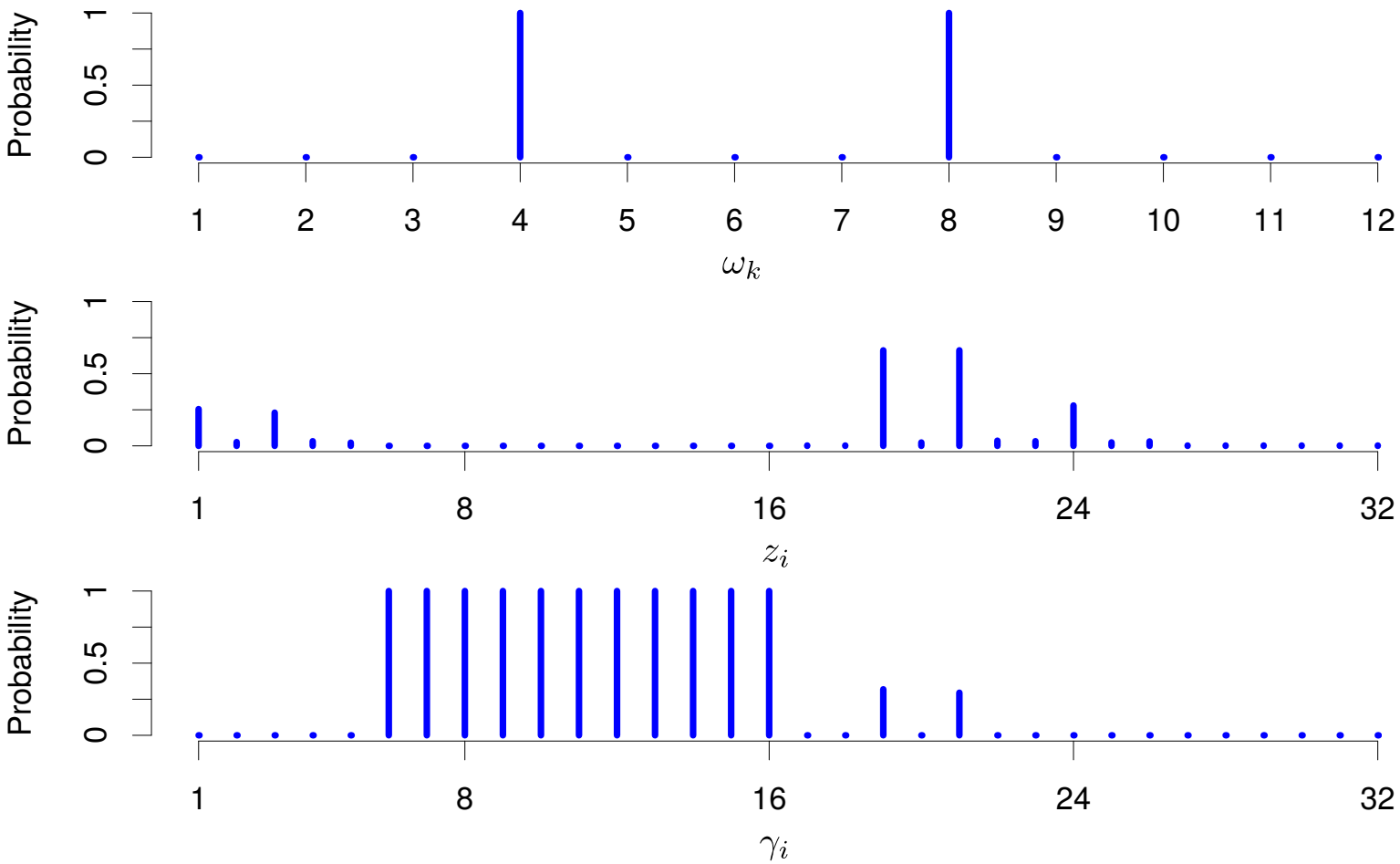
$N = 200$, $d = 2,000$ and we use for $\mathbf{W}$ the pattern earlier shown. $K = 12$, $\sigma^2_{(k)} = 0.5$, $\forall k$ and $w_i^{(k)} \sim \text{Student}(\text{df} = 5)$.

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	$2 \cdot 10^3 \pm 4 \cdot 10^2$
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.

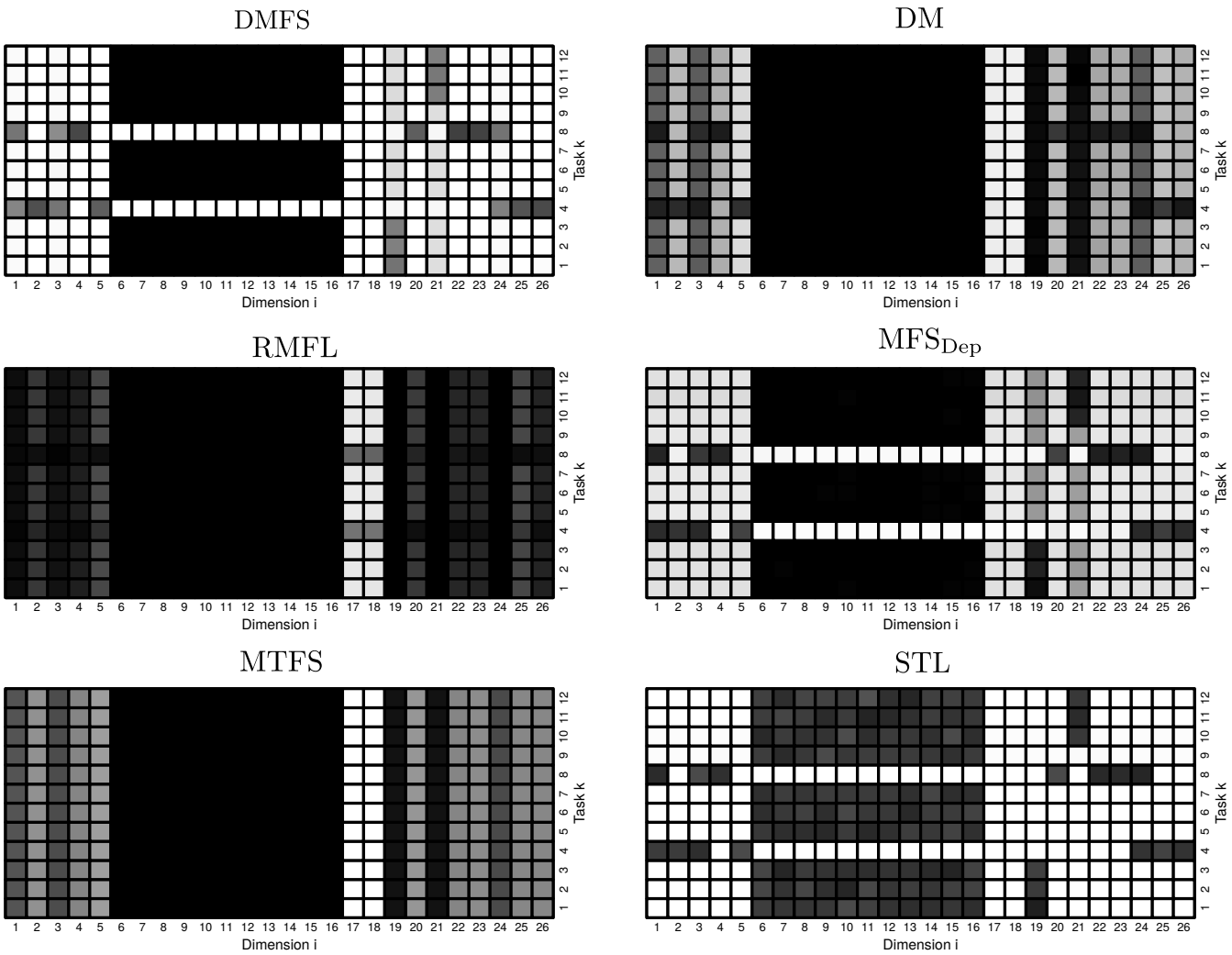
Experiments with synthetic data (II)

Average posterior probabilities of the binary latent variables.



Experiments with synthetic data (III)

Average probability that each coefficient is different from zero.

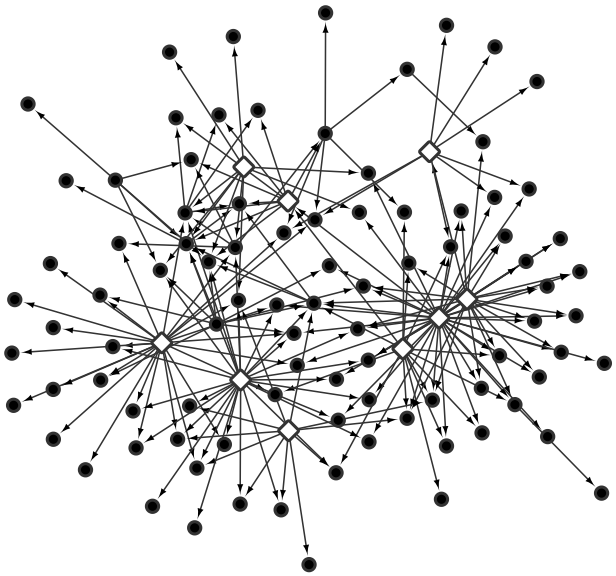


DM and RMFL **shrink** relevant coefficients (Hernández-Lobato et al., 2013).

Reconstruction of gene regulatory networks

$\mathbf{X}$ contains N measurements of log mRNA: $\mathbf{X} \approx \mathbf{X}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{E}$. To estimate $\mathbf{W}$ we create d tasks where $\mathbf{X}^{(k)}$ is given by $\mathbf{X}$ with column k set to 0. The targets are the elements in that column.

DREAM 4 in silico challenge: we use GeneNetWeaver to generate 100 networks with 100 genes and 90 measurements. The edge $j \to i$ is predicted when $p(w_i^{(j)} \neq 0)$ exceeds $\zeta \in [0, 1]$.



Method	AUROC
MFS	0.73±0.05
DMFS	**0.84±0.05**
DM	0.76±0.06
$\mathrm{MFS}_{\mathrm{Dep}}$	0.79±0.06
RMFL	0.79±0.05
STL	0.72±0.04

Average AUROC of the winning solution of the challenge: 0.75.

Denoising of natural images (I)

We consider denoising the 256×256 *house* image when it has been contaminated with Gaussian noise with $\sigma_{(k)}$ equal to 25, 50 and 75.

We generate $62,001$ blocks of 8×8 pixels and let each block be a different task. We consider 64 groups of non-overlapping blocks.

$\mathbf{y}^{(k)} = \mathbf{X}^{(k)}\mathbf{w}^{(k)} + \boldsymbol{\epsilon}^{(k)}$, where $\mathbf{X}^{(k)}$ is a Haar wavelet orthonormal basis.

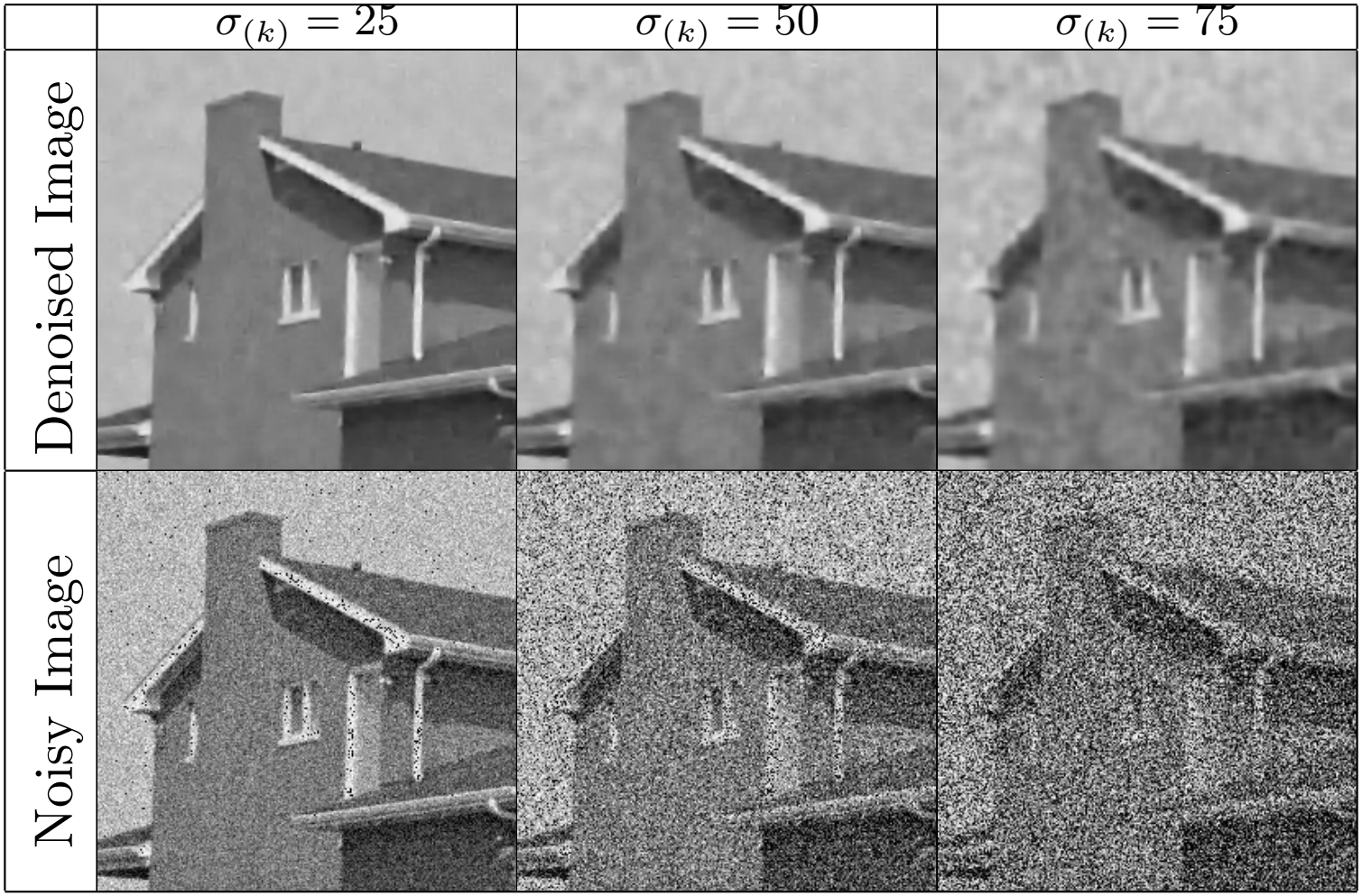
Peak-to-signal ratio for each method.

Method	$\sigma_{(k)} = 25$	$\sigma_{(k)} = 50$	$\sigma_{(k)} = 75$
MFS	25.89	23.89	23.87
DMFS	**30.67**	**27.21**	**25.23**
DM	28.50	25.91	24.24
MFS_{Dep}	30.46	25.74	23.65
RMFL	28.35	25.56	24.09
STL	30.55	26.26	23.40

In these experiments $N = 64$, $d = 64$.

Denoising of natural images (II)

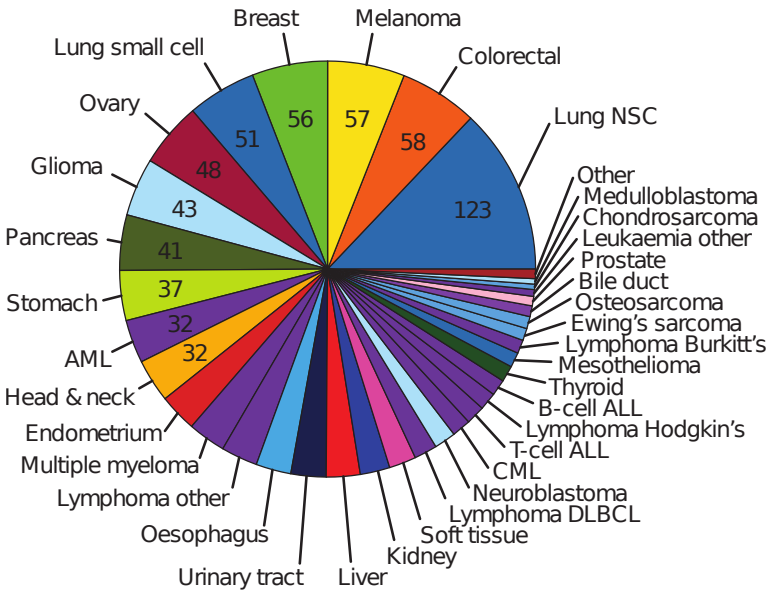
Denoising results for the proposed method, DMFS.



Anti-cancer drug sensitivity prediction (I)

The dataset of Barretina et al. (2012) has mRNA levels for 294 cell lines and their sensitivity level to 24 anti-cancer drugs.

We consider only the 1,000 genes with the largest IQR distance.



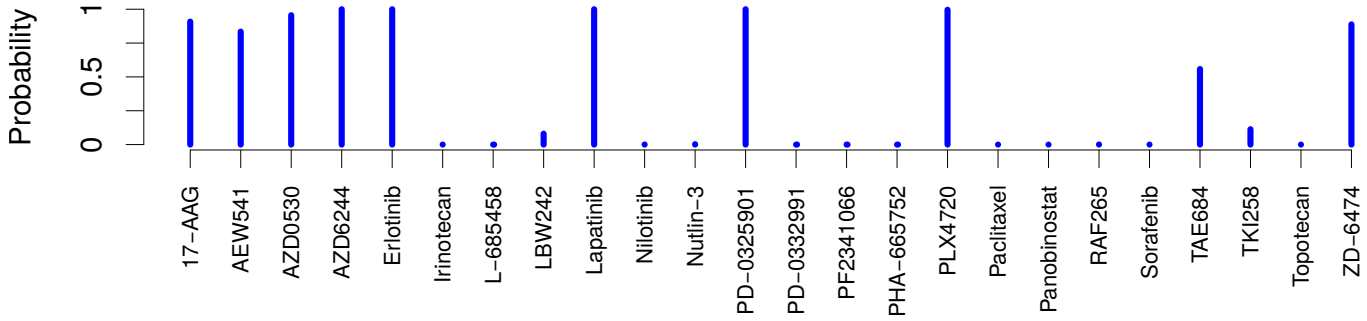
Method	RMSE
DMFS	0.715±0.050
DM	**0.703±0.050**
RMFL	**0.703±0.050**
MFS_{Dep}	0.704±0.051
MFS	0.734±0.053
STL	0.743±0.051

Extracted from (Barretina et al., 2012)

The differences with respect to STL and MFS are statistically significant. DM reduces in these experiments to the **group LASSO**.

Anti-cancer drug sensitivity prediction (II)

We compare DMFS and RMFL to identify outlier tasks.



Avg. prob. that each drug (task) is an outlier, as estimated by DMFS.

We evaluate the group LASSO on the non-outlier tasks (identified by DMFS or RMFL) when outlier tasks are thrown away and when they are not.

RMSE of the group LASSO in the non-outlier tasks

	Outlier Tasks Not Removed	**Outlier Tasks Removed**	**Improvement in RMSE** $\times 10^{-3}$
DMFS	0.672±0.070	0.668±0.072	3.64±1.25
RMFL	0.686±0.069	0.684±0.074	1.97±1.22

The first improvement is **statistically significant**. The second is not.

Conclusions

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

References I

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

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.

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

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*, volume 6321, 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. *Journal of Machine Learning Research*, 14:1891–1945, 2013.

References II

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 Titterington, D.M. Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367:4237, 2009.

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.

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

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 *International Conference on Data Mining*, pp. 332–342, 2007.