Supplementary Material for: A Probabilistic Model for Dirty Multi-task Feature Selection

Daniel Hernández-Lobato
Universidad Autónoma de Madrid
Francisco Tomás y Valiente 11
28049, Madrid, Spain
daniel.hernandez@uam.com

José Miguel Hernández-Lobato
Harvard University
33 Oxford street
Cambridge, MA 02138, USA
jmhl@seas.harvard.edu

Zoubin Ghahramni
University of Cambridge
Trumpington Street
Cambridge CB2 1PZ, UK
zoubin@eng.cam.ac.uk

1 Expectation Propagation

In this section we give all the necessary details to implement the EP algorithm for the proposed method in the main manuscript, *i.e.* DMTFS. We describe how to compute the EP posterior approximation from the product of all approximate factors and how to implement the EP updates to refine each approximate factor. Finally, although not used in the experiments, we also give an intuitive idea about how to compute the EP approximation to the marginal likelihood. Recall from the main manuscript that the approximate factors replace the corresponding factors in the joint distribution $p(\mathbf{Y}, \mathbf{W}, \mathbf{\Omega}, \boldsymbol{\rho}, \boldsymbol{\sigma}^2 | \boldsymbol{\mathcal{X}})$. The resulting approximate joint distribution is then normalized to get the EP posterior approximation, and the normalization constant is the approximation to the marginal likelihood.

1.1 Natural Parameter Representation

To make easier the description of the EP algorithm, we will work with the natural parameters of each approximate factor. We have in consequence that the n-th likelihood factor of task k, *i.e.*, $p(y_n^{(k)} | \mathbf{w}^{(k)}, \mathbf{x}_n^{(k)}, \sigma_{(k)}^2) = \mathcal{N}(y_n^{(k)} | (\mathbf{w}^{(k)})^{\mathrm{T}} \mathbf{x}_n^{(k)}, \sigma_{(k)}^2)$, is in practice approximated as:

$$p(y_n^{(k)} | \mathbf{w}^{(k)}, \mathbf{x}_n^{(k)}, \sigma_{(k)}^2) \approx \tilde{f}_n^{(k)} \left(\mathbf{w}^{(k)}, \sigma_{(k)}^2 \right) = \tilde{c}_n^{(k)} \exp \left\{ \tilde{m}_n^{(k)} (\mathbf{w}^{(k)})^{\mathrm{T}} \mathbf{x}_n^{(k)} - \right.$$
$$\left. - \frac{\tilde{v}_n^{(k)}}{2} (\mathbf{w}^{(k)})^{\mathrm{T}} \mathbf{x}_n^{(k)} (\mathbf{x}_n^{(k)})^{\mathrm{T}} \mathbf{w}^{(k)} \right\} \times \exp \left\{ -\tilde{a}_n^{(k)} \log(\sigma_{(k)}^2) - \frac{\tilde{b}_n^{(k)}}{\sigma_{(k)}^2} \right\} , \quad (1)$$

where the natural parameters $\tilde{c}_n^{(k)}$, $\tilde{m}_n^{(k)}$, $\tilde{v}_n^{(k)}$, $\tilde{a}_n^{(k)}$ and $\tilde{b}_n^{(k)}$ are to be fixed by EP. Note that the first exponential function corresponds to an un-normalized Gaussian distribution and that the second exponential function corresponds to an un-normalized Inverse Gamma distribution. Thus, the approximation is the same as the one described in the main manuscript, but with a different parameter representation.

The approximation for each factor $p(w_i^{(k)} | \mathbf{\Omega})$ corresponding to the robust prior distribution for

the i-th coefficient of the k-th tasks, $w_i^{(k)}$, is:

$$p(w_i^{(k)}|\mathbf{\Omega}) \approx \tilde{g}_i^{(k)}\left(w_i^{(k)}, z_i, \omega_k, \gamma_i, \tau_i^{(k)}, \eta_i^{(k)}\right) = \tilde{s}_i^{(k)} \exp\left\{\tilde{m}_i^{(k)} w_i^{(k)} - \frac{\tilde{v}_i^{(k)}}{2}\left(w_i^{(k)}\right)^2\right\} \times$$
$$\exp\left\{z_i \tilde{p}_z^{(i,k)}\right\} \exp\left\{\omega_k \tilde{p}_\omega^{(i,k)}\right\} \exp\left\{\gamma_i \tilde{p}_\gamma^{(i,k)}\right\} \exp\left\{\tau_i^{(k)} \tilde{p}_\tau^{(i,k)}\right\} \exp\left\{\eta_i^{(k)} \tilde{p}_\eta^{(i,k)}\right\}. \quad (2)$$

Again, the parameters $\tilde{c}_i^{(k)}$, $\tilde{m}_i^{(k)}$, $\tilde{v}_i^{(k)}$, $\tilde{p}_z^{(i,k)}$, $\tilde{p}_\omega^{(i,k)}$, $\tilde{p}_\gamma^{(i,k)}$, $\tilde{p}_\tau^{(i,k)}$ and $\tilde{p}_\eta^{(i,k)}$ are natural parameters to be fixed by EP. Additionally, the first exponential function is again an un-normalized Gaussian distribution and $\exp\{z_i \tilde{p}_z^{(i,k)}\}$ is an un-normalized Bernoulli distribution for the binary random variable z_i (the same applies for the other random binary variables $g_i^{(k)}$ depends on). Thus, the approximation is again the same as the one described in the main manuscript, but with a different parameter representation.

The approximation of each factor corresponding to the prior for the binary latent variables is:

$$p(z_i|\rho_z) \approx \tilde{h}_z^{(i)}(z_i, \rho_z) = \tilde{k}_z^{(i)} \exp\left\{z_i \tilde{p}_z^{(i)}\right\} \exp\left\{\tilde{a}_z^{(i)} \log(\rho_z) + \tilde{b}_z^{(i)} \log(1-\rho_z)\right\} \quad \forall i\,, \quad (3)$$

$$p(\omega_k|\rho_\omega) \approx \tilde{h}_\omega^{(k)}(\omega_k, \rho_\omega) = \tilde{k}_\omega^{(k)} \exp\left\{\omega_k \tilde{p}_\omega^{(k)}\right\} \exp\left\{\tilde{a}_\omega^{(k)} \log(\rho_\omega) + \tilde{b}_\omega^{(k)} \log(1-\rho_\omega)\right\} \quad \forall k\,, \quad (4)$$

$$p(\gamma_i|\rho_\gamma) \approx \tilde{h}_\gamma^{(i)}(\gamma_i, \rho_\gamma) = \tilde{k}_\gamma^{(i)} \exp\left\{\gamma_i \tilde{p}_\gamma^{(i)}\right\} \exp\left\{\tilde{a}_\gamma^{(i)} \log(\rho_\gamma) + \tilde{b}_\gamma^{(i)} \log(1-\rho_\gamma)\right\} \quad \forall i\,, \quad (5)$$

$$p(\tau_i^{(k)}|\rho_\tau) \approx \tilde{h}_\tau^{(i,k)}(\tau_i^{(k)}, \rho_\tau) = \tilde{k}_\tau^{(i,k)} \exp\left\{\tau_i^{(k)} \tilde{p}_\tau^{(i,k)}\right\}$$
$$\exp\left\{\tilde{a}_\tau^{(i,k)} \log(\rho_\tau) + \tilde{b}_\tau^{(i,k)} \log(1-\rho_\tau)\right\} \quad \forall i,k\,, \quad (6)$$

$$p(\eta_i^{(k)}|\rho_\eta) \approx \tilde{h}_\eta^{(i,k)}(\eta_i^{(k)}, \rho_\eta) = \tilde{k}_\eta^{(i,k)} \exp\left\{\eta_i^{(k)} \tilde{p}_\eta^{(i,k)}\right\}$$
$$\exp\left\{\tilde{a}_\eta^{(i,k)} \log(\rho_\eta) + \tilde{b}_\eta^{(i,k)} \log(1-\rho_\eta)\right\} \quad \forall i,k\,, \quad (7)$$

where all parameters with the superscript $\tilde{}$ are natural parameters to be fixed by EP. In this case, each factor is the product of an un-normalized Bernoulli distribution and an un-normalized Beta distribution, expressed in terms of their natural parameters.

We also show how to express in this notation the factors that need not be approximated. These include the factors corresponding to the priors for each $\sigma_{(k)}^2$ and the factors corresponding to the priors for each activation probability p_z, p_ω, p_γ, p_τ and p_η. In particular,

$$p(\sigma_{(k)}^2) = \text{InvGam}(\sigma_{(k)}^2|5,5) = \frac{5^5}{\Gamma(5)} \exp\left\{-(5+1)\log(\sigma_{(k)}^2) - \frac{5}{\sigma_{(k)}^2}\right\} \quad \forall k\,, \quad (8)$$

$$p(\rho_z) = \text{Beta}(\rho_z|1,1) = \frac{1}{\beta(1,1)} \exp\left\{0 \cdot \log(\rho_z) + 0 \cdot \log(1-\rho_z)\right\} \quad (9)$$

$$p(\rho_\omega) = \text{Beta}(\rho_\omega|1,1) = \frac{1}{\beta(1,1)} \exp\left\{0 \cdot \log(\rho_\omega) + 0 \cdot \log(1-\rho_\omega)\right\}, \quad (10)$$

$$p(\rho_\gamma) = \text{Beta}(\rho_\gamma|1,1) = \frac{1}{\beta(1,1)} \exp\left\{0 \cdot \log(\rho_\gamma) + 0 \cdot \log(1-\rho_\gamma)\right\}, \quad (11)$$

$$p(\rho_\tau) = \text{Beta}(\rho_\tau|1,1) = \frac{1}{\beta(1,1)} \exp\left\{0 \cdot \log(\rho_\tau) + 0 \cdot \log(1-\rho_\tau)\right\}, \quad (12)$$

$$p(\rho_\eta) = \text{Beta}(\rho_\eta|1,1) = \frac{1}{\beta(1,1)} \exp\left\{0 \cdot \log(\rho_\eta) + 0 \cdot \log(1-\rho_\eta)\right\}, \quad (13)$$

where $\Gamma(\cdot)$ is the gamma function and $\beta(\cdot,\cdot)$ is the beta function.

1.2 Computation of the EP Posterior Approximation

In this section we show how to compute the posterior approximation given the approximate factors. For this, we use the joint approximation $\tilde{q}$ which is defined as the product of all approximate factors

and the factors that need not be approximated. Assume K learning tasks with N_k examples and d features each. Then,

$$\begin{aligned}\tilde{q}(\mathbf{W}, \boldsymbol{\Omega}, \boldsymbol{\rho}, \boldsymbol{\sigma}^2) = & \left[\prod_{k=1}^{K}\prod_{n=1}^{N_k} \tilde{f}_n^{(k)}(\mathbf{w}^{(k)}, \sigma_{(k)}^2)\right]\left[\prod_{i=1}^{d}\prod_{k=1}^{K} \tilde{g}_i^{(k)}(w_i^{(k)}, z_i, \omega_k, \gamma_i, \tau_i^{(k)}, \eta_i^{(k)})\right] \times \\ & \times \left[\prod_{i=1}^{d} \tilde{h}_z^{(i)}(z_i, \rho_z)\right]\left[\prod_{k=1}^{K} \tilde{h}_\omega^{(k)}(\omega_k, \rho_\omega)\right]\left[\prod_{i=1}^{d} \tilde{h}_\gamma^{(i)}(\gamma_i, \rho_\gamma)\right]\left[\prod_{i=1}^{d}\prod_{k=1}^{K} \tilde{h}_\tau^{(i,k)}(\tau_i^{(k)}, \rho_\tau)\right] \\ & \times \left[\prod_{i=1}^{d}\prod_{k=1}^{K} \tilde{h}_\eta^{(i,k)}(\eta_i^{(k)}, \rho_\eta)\right]\left[\prod_{k=1}^{K} p(\sigma_{(k)}^2)\right] p(\rho_z)p(\rho_\omega)p(\rho_\gamma)p(\rho_\tau)p(\rho_\eta)\,, \end{aligned} \tag{14}$$

which is an un-normalized distribution inside the exponential family $\mathcal{F}$ defined in the main manuscript. After $\tilde{q}$ is normalized to integrate to one, the EP posterior approximation is obtained. In particular,

$$\begin{aligned} q(\mathbf{W}, \boldsymbol{\Omega}, \boldsymbol{\rho}, \boldsymbol{\sigma}^2) &= \frac{\tilde{q}(\mathbf{W}, \boldsymbol{\Omega}, \boldsymbol{\rho}, \boldsymbol{\sigma}^2)}{Z_q}\,, \\ &= \left[\prod_{k=1}^{K} \mathcal{N}(\mathbf{w}^{(k)}|\mathbf{m}^{(k)}, \mathbf{V}^{(k)})\right]\left[\prod_{i=1}^{d} \text{Bern}(z_i|p_z^{(i)})\right]\left[\prod_{k=1}^{K} \text{Bern}(\omega_k|p_\omega^{(k)})\right] \times \\ &\times \left[\prod_{i=1}^{d} \text{Bern}(\gamma_i|p_\gamma^{(i)})\right]\left[\prod_{i=1}^{d}\prod_{k=1}^{K} \text{Bern}(\tau_i^{(k)}|p_\tau^{(i,k)})\right]\left[\prod_{i=1}^{d}\prod_{k=1}^{K} \text{Bern}(\eta_i^{(k)}|p_\eta^{(i,k)})\right] \times \\ &\times \left[\prod_{k=1}^{K} \text{InvGam}(\sigma_{(k)}^2|a_k, b_k)\right] \text{Beta}(\rho_z|a_z, b_z)\text{Beta}(\rho_\omega|a_\omega, b_\omega)\times \\ &\times \text{Beta}(\rho_\gamma|a_\gamma, b_\gamma)\text{Beta}(\rho_\tau|a_\tau, b_\tau)\text{Beta}(\rho_\eta|a_\eta, b_\eta) \end{aligned} \tag{15}$$

where Z_q is the normalization constant of $\tilde{q}$, which can be used to approximate the marginal likelihood of the model. The parameters of the posterior approximation q such as $\mathbf{m}^{(k)}$, $\mathbf{V}^{(k)}$, etc. can be computed by summing the natural parameters of the factors in $\tilde{q}$. This will give the natural parameters of q. Given these natural parameters, the corresponding standard parameters can be easily obtained. Let the superscript $\hat{}$ denote the corresponding natural parameters. For example, in the case of the multi-variate Gaussian distribution, the natural parameters $\hat{\mathbf{m}}^{(k)}$ and $\hat{\mathbf{V}}^{(k)}$ are defined as $\hat{\mathbf{m}}^{(k)} = (\mathbf{V}^{(k)})^{-1}\mathbf{m}^{(k)}$ and $\hat{\mathbf{V}}^{(k)} = (\mathbf{V}^{(k)})^{-1}$. By multiplying all the approximate factors that depend on $\mathbf{w}^{(k)}$, we have that

$$\begin{aligned} \hat{\mathbf{V}}^{(k)} &= \left((\mathbf{X}^{(k)})^{\text{T}}\boldsymbol{\Delta}_k\mathbf{X}^{(k)} + \boldsymbol{\Pi}_k\right)\,, & \mathbf{V}_k &= (\hat{\mathbf{V}}^{(k)})^{-1}\,, \\ \hat{\mathbf{m}}^{(k)} &= (\mathbf{X}^{(k)})^{\text{T}}\tilde{\boldsymbol{\upsilon}}^{(k)} + \tilde{\mathbf{m}}^{(k)}\,, & \mathbf{m}^{(k)} &= \mathbf{V}^{(k)}\hat{\mathbf{m}}^{(k)}\,, \end{aligned} \tag{16}$$

where $\boldsymbol{\Delta}_k$ is a diagonal matrix of size $N_k \times N_k$ whose diagonal elements are given by $\tilde{v}_n^{(k)}$, the parameter of $\tilde{f}_k$; $\boldsymbol{\Pi}_k$ is a $d \times d$ diagonal matrix whose elements are given by $\tilde{v}_i^{(k)}$, the parameter of $\tilde{g}_i^{(k)}$; $\tilde{\mathbf{v}}^{(k)}$ is a N_k dimensional vector with components equal to $\tilde{v}_n^{(k)}$, the parameter of $\tilde{f}_k$; $\tilde{\boldsymbol{\upsilon}}^{(k)}$ is a N_k dimensional vector with components equal to $\tilde{m}_n^{(k)}$, the parameter of $\tilde{f}_k$; and $\tilde{\mathbf{m}}^{(k)}$ is a d dimensional vector with components equal to $\tilde{m}_i^{(k)}$, the parameter of $\tilde{g}_i^{(k)}$. This is valid for each k. Furthermore, the inversion of $\hat{\mathbf{V}}^{(k)}$ can be carried out with cost $\mathcal{O}(N_k^2 d)$ using the Woodbury

formula. The parameters of the other distributions in q are computed similarly. In particular,

$$\begin{aligned}
\hat{p}_z^{(i)} &= \sum_{k=1}^{K} \tilde{p}_z^{(i,k)} + \tilde{p}_z^{(i)}\,, & p_z^{(i)} &= \frac{1}{1+\exp(-\hat{p}_z^{(i)})}\,,\\
\hat{p}_\omega^{(k)} &= \sum_{i=1}^{d} \tilde{p}_\omega^{(i,k)} + \tilde{p}_\omega^{(k)}\,, & p_\omega^{(k)} &= \frac{1}{1+\exp(-\hat{p}_\omega^{(k)})}\,,\\
\hat{p}_\gamma^{(i)} &= \sum_{k=1}^{K} \tilde{p}_\gamma^{(i,k)} + \tilde{p}_\gamma^{(i)}\,, & p_\gamma^{(i)} &= \frac{1}{1+\exp(-\hat{p}_\gamma^{(i)})}\,,\\
\hat{p}_\tau^{(i,k)} &= \tilde{p}_\tau^{(i,k)} + \tilde{\tilde{p}}_\tau^{(i,k)}\,, & p_\tau^{(i,k)} &= \frac{1}{1+\exp(-\hat{p}_\tau^{(i,k)})}\,,\\
\hat{p}_\eta^{(i,k)} &= \tilde{p}_\eta^{(i,k)} + \tilde{\tilde{p}}_\eta^{(i,k)}\,, & p_\eta^{(i,k)} &= \frac{1}{1+\exp(-\hat{p}_\eta^{(i,k)})}\,,
\end{aligned} \tag{17}$$

where $\tilde{p}_z^{(i,k)}$, $\tilde{p}_\omega^{(i,k)}$, $\tilde{p}_\gamma^{(i,k)}$, $\tilde{p}_\tau^{(i,k)}$ and $\tilde{p}_\eta^{(i,k)}$ are the parameters of $\tilde{g}_i^{(k)}$, and $\tilde{p}_z^{(i)}$, $\tilde{p}_\omega^{(k)}$, $\tilde{p}_\gamma^{(i)}$, $\tilde{\tilde{p}}_\tau^{(i,k)}$ and $\tilde{\tilde{p}}_\eta^{(i,k)}$ are the parameters of $h_z^{(i)}$, $h_\omega^{(k)}$, $h_\gamma^{(i)}$, $h_\tau^{(i,k)}$ and $h_\eta^{(i,k)}$, respectively. The parameters of the Inverse Gammas are:

$$\begin{aligned}
\hat{a}_k &= \sum_{n=1}^{N_k} \tilde{a}_n^{(k)} + 5 + 1\,, & a_k &= \hat{a}_k - 1\,, \quad \forall k\,,\\
\hat{b}_k &= \sum_{n=1}^{N_k} \tilde{b}_n^{(k)} + 5 + 1\,, & b_k &= \hat{b}_k - 1\,, \quad \forall k\,.
\end{aligned} \tag{18}$$

If the same noise level is assumed for each task these computations are:

$$\begin{aligned}
\hat{a}_k &= \sum_{k=1}^{K}\sum_{n=1}^{N_k} \tilde{a}_n^{(k)} + 5 + 1\,, & a_k &= \hat{a}_k - 1\,, \quad \forall k\,,\\
\hat{b}_k &= \sum_{k=1}^{K}\sum_{n=1}^{N_k} \tilde{b}_n^{(k)} + 5 + 1\,, & b_k &= \hat{b}_k - 1\,, \quad \forall k\,.
\end{aligned} \tag{19}$$

Finally, the parameters of the beta distributions are:

$$\begin{aligned}
\hat{a}_z &= \sum_{i=1}^{d} \tilde{a}_z^{(i)}\,, & a_z &= \hat{a}_z + 1\,, & \hat{b}_z &= \sum_{i=1}^{d} \tilde{b}_z^{(i)}\,, & b_z &= \hat{b}_z + 1\,,\\
\hat{a}_\omega &= \sum_{k=1}^{K} \tilde{a}_\omega^{(k)}\,, & a_\omega &= \hat{a}_\omega + 1\,, & \hat{b}_\omega &= \sum_{k=1}^{K} \tilde{b}_\omega^{(k)}\,, & b_\omega &= \hat{b}_\omega + 1\,,\\
\hat{a}_\gamma &= \sum_{i=1}^{d} \tilde{a}_\gamma^{(i)}\,, & a_\gamma &= \hat{a}_\gamma + 1\,, & \hat{b}_\gamma &= \sum_{i=1}^{d} \tilde{b}_\gamma^{(i)}\,, & b_\gamma &= \hat{b}_\gamma + 1\,,\\
\hat{a}_\tau &= \sum_{k=1}^{K}\sum_{i=1}^{d} \tilde{a}_\tau^{(i,k)}\,, & a_\tau &= \hat{a}_\tau + 1\,, & \hat{b}_\tau &= \sum_{k=1}^{K}\sum_{i=1}^{d} \tilde{b}_\tau^{(i,k)}\,, & b_\tau &= \hat{b}_\tau + 1\,,\\
\hat{a}_\eta &= \sum_{k=1}^{K}\sum_{i=1}^{d} \tilde{a}_\eta^{(i,k)}\,, & a_\eta &= \hat{a}_\eta + 1\,, & \hat{b}_\eta &= \sum_{k=1}^{K}\sum_{i=1}^{d} \tilde{b}_\eta^{(i,k)}\,, & b_\eta &= \hat{b}_\eta + 1\,,
\end{aligned} \tag{20}$$

where the parameters $\tilde{a}_z^{(i)}$, $\tilde{a}_\omega^{(k)}$, $\tilde{a}_\gamma^{(i)}$, $\tilde{a}_\tau^{(i,k)}$, $\tilde{a}_\eta^{(i,k)}$ and $\tilde{b}_z^{(i)}$, $\tilde{b}_\omega^{(k)}$, $\tilde{b}_\gamma^{(i)}$, $\tilde{b}_\tau^{(i,k)}$, $\tilde{b}_\eta^{(i,k)}$ are the parameters of the approximate factors $\tilde{h}_z^{(i)}(z_i, \rho_z)$, $\tilde{h}_\omega^{(k)}(\omega_k, \rho_\omega)$, $\tilde{h}_\gamma^{(i)}(\gamma_i, \rho_\gamma)$, $\tilde{h}_\tau^{(i,k)}(\tau_i^{(k)}, \rho_\tau)$ and $\tilde{h}_\eta^{(i,k)}(\eta_i^{(k)}, \rho_\eta)$.

1.3 Update of the Approximate Factors

In this section we give specific details of how to update the parameters of each of the approximate factors. For that, we will assume that there is an old distribution q^{old} that has been computed by removing the corresponding approximate factor $\tilde{f}$ from the posterior q. Namely, $q^{\text{old}} = q/\tilde{f}$. The parameters of q^{old} are exactly the same parameters as those of q, but where we have subtracted from the natural parameters of q the natural parameters of the approximate factor $\tilde{f}$. The updated factor is simply given, up to a multiplicative constant, by the ratio $q^{\text{new}}/q^{\text{old}}$, where q^{new} is the posterior approximation that minimizes $\text{KL}(fq^{\text{old}}||q^{\text{new}})$ and f is the corresponding exact factor. Let Z_f be the normalization constant of fq^{old}. As described in the main manuscript q^{new} is obtained by setting its moments to those of fq^{old}. Furthermore, these moments can be obtained from the derivatives of $\log Z_f$ with respect to the natural parameters of q^{old} [6]. Thus, the only quantity that is needed to perform the EP updates is Z_f. We explain how to compute its value for each approximate factor.

In our EP implementation we update all factors in parallel as in [8]. We also employ damped updates. When the EP updates are damped the parameters of the approximate factors cannot change too much. The practical implementation of damping is trivial and is described in detail in [2]. In the rest of the section we annotate with the superscript $^-$ to the corresponding parameters of q^{old}. For example, the parameter $m_i^{(k)}$ that describes the posterior mean of $w_i^{(k)}$ under q will be denoted by $\overline{m}_i^{(k)}$ in the case that it is actually the posterior mean of that variable under q^{old}. Typically, each approximate factor will only depend on a few of the latent variables in q. This means that we can safely marginalize all other variables in q before computing q^{old}.

To simplify the EP updates, we will in general work with the natural parameters of q^{old}. These are obtained simply by subtracting from the natural parameters of q the natural parameters of the corresponding approximate factor. From now on, we add the supper script $\hat{}$ to the standard parameters of q^{old} to denote natural parameters. Continuing with the previous example, the first natural parameter of the Gaussian posterior distribution of $w_i^{(k)}$ under q^{old} is $\hat{\overline{m}}_i^{(k)} = \overline{m}_i^{(k)}/\overline{v}_i^{(k)}$, where $\overline{v}_i^{(k)}$ is the variance of $w_i^{(k)}$ under q^{old} and $\overline{m}_i^{(k)}$ is its mean. The second natural parameter is $\hat{\overline{v}}_i^{(k)} = 1/\overline{v}_i^{(k)}$. However, as described before, it is better in practice to compute the natural parameters of q^{old} by subtracting the natural parameters of the corresponding approximate factor to the natural parameters of q.

1.3.1 Approximate Factors Corresponding to the Likelihood

Consider the exact factor $f_n^{(k)}((\mathbf{w}^{(k)})^{\text{T}}\mathbf{x}_n^{(k)}, \sigma_{(k)}^2) = \mathcal{N}(y_n^{(k)}|(\mathbf{w}^{(k)})^{\text{T}}\mathbf{x}_n^{(k)}, \sigma_{(k)}^2)$. To update the corresponding approximate factor $\tilde{f}_n^{(k)}$, we define the variable $b_n^{(k)} = (\mathbf{w}^{(k)})^{\text{T}}\mathbf{x}_n^{(k)}$. The mean and variance of $b_n^{(k)}$ under q can be obtained from the standard parameters of this distribution. They are $\mu_n^{(k)} = (\mathbf{m}^{(k)})^{\text{T}}\mathbf{x}_n^{(n)}$ and $\Sigma_n^{(k)} = (\mathbf{x}_n^{(k)})^{\text{T}}\mathbf{V}^{(k)}\mathbf{x}_n^{(k)}$, respectively. They can be efficiently computed if the Woodbury formula is used to evaluate $\mathbf{V}^{(k)}$. The q^{old} distribution is then given as:

$$q^{\text{old}}(b_n^{(k)}, \sigma_{(k)}^2) = \exp\left\{ b_n^{(k)}\hat{\overline{\mu}}_n^{(k)} - \frac{\hat{\overline{\Sigma}}_n^{(k)}}{2}(b_n^{(k)})^2 \right\} \exp\left\{ -\hat{\overline{a}}_k \log(\sigma_{(k)}^2) - \frac{\hat{\overline{b}}_k}{\sigma_{(k)}^2} \right\}, \tag{21}$$

where $\hat{\overline{\mu}}_n^{(k)} = \mu_n^{(k)}/\Sigma_n^{(k)} - \tilde{m}_n^{(k)}$, $\hat{\overline{\Sigma}}_n^{(k)} = 1/\Sigma_n^{(k)} - \tilde{v}_n^{(k)}$, $\hat{\overline{a}}_k = \hat{a}_k - \tilde{a}_n^{(k)}$ and $\hat{\overline{b}}_k = \hat{b}_k - \tilde{b}_n^{(k)}$. In these last expressions $\tilde{m}_n^{(k)}$, $\tilde{v}_n^{(k)}$, $\tilde{a}_n^{(k)}$ and $\tilde{b}_n^{(k)}$ are the natural parameters of the approximate factor $\tilde{f}_n^{(n)}$. The other variables are the parameters of q. We are now interested in computing the normalization

constant of $f_n^{(k)} q^{\text{old}}$, $Z_{f_n^{(k)}}$. In particular,

$$
\begin{aligned}
Z_{f_n^{(k)}} &= \int f_k^{(k)}(b_n^{(k)}, \sigma_{(k)}^2) q^{\text{old}}(b^{(k)}, \sigma_{(k)}^2) db_n^{(k)} d\sigma_{(k)}^2 \\
&= \int \mathcal{N}(y_n^{(k)} | b_n^{(k)}, \sigma_{(k)}^2) q^{\text{old}}(b^{(k)}, \sigma_{(k)}^2) db_n^{(k)} d\sigma_{(k)}^2 \\
&= \frac{\Gamma(\hat{\bar{a}}_k - 1)}{(\hat{\bar{b}}_k)^{\hat{\bar{a}}_k - 1}} \int \mathcal{T}(y_n^{(k)} | b_n^{(k)}, s_n^{(k)}, \nu_n^{(k)}) \exp\left\{ b_n^{(k)} \hat{\bar{\mu}}_n^{(k)} - \frac{\hat{\bar{\Sigma}}_n^{(k)}}{2} (b_n^{(k)})^2 \right\} db_n^{(k)} \\
&\approx \frac{\Gamma(\hat{\bar{a}}_k - 1)}{(\hat{\bar{b}}_k)^{\bar{a}_k - 1}} \int \mathcal{N}(y_n^{(k)} | b_n^{(k)}, \text{Var}_n^{(k)}) \exp\left\{ b_n^{(k)} \hat{\bar{\mu}}_n^{(k)} - \frac{\hat{\bar{\Sigma}}_n^{(k)}}{2} (b_n^{(k)})^2 \right\} db_n^{(k)} \\
&= \frac{\Gamma(\hat{\bar{a}}_k - 1)}{(\hat{\bar{b}}_k)^{\hat{\bar{a}}_k - 1}} \mathcal{N}\left(y_n^{(k)} | \hat{\bar{\mu}}_n^{(k)} / \hat{\bar{v}}_n^{(k)}, \text{Var}_n^{(k)} + (\hat{\bar{v}}_n^{(k)})^{-1} \right) \sqrt{\frac{2\pi}{\hat{\bar{v}}_n^{(k)}}} \exp\left\{ \frac{1}{2} \frac{(\hat{\bar{\nu}}_n^{(k)})^2}{\hat{\bar{v}}_n^{(k)}} \right\}, \qquad (22)
\end{aligned}
$$

where $\mathcal{T}(\cdot | m, s, \nu)$ denotes a Student's T distribution with location parameter m, scale parameter s and number of degrees of freedom ν; $\Gamma(\cdot)$ is the gamma function;$\nu_n^{(k)} = 2(\hat{\bar{a}}_n - 1)$, $\text{Var}_n^{(k)} = \hat{\bar{b}}_k / (\hat{\bar{a}}_k - 2)$, and $s_n^{(k)} = \hat{\bar{b}}_k / (\hat{\bar{a}}_k - 1)$. Note that we have approximated the Student's T distribution with a Gaussian with the same mean and the same variance. This approximation is accurate for high values of the degrees of freedom $\nu_n^{(k)}$ and has also been employed in [4, 3] to carry out approximate inference.

The moments of $f_n^{(k)} q^{\text{old}}$ required to find q^{new} are all obtained from $Z_{f_n^{(k)}}$ [6]. In particular, these moments are:

$$
\begin{aligned}
\mathbb{E}_{f_n^{(k)} q^{\text{old}}}\left[b_k^{(n)} \right] &\approx \left. \frac{\partial \log Z_{f_n^{(k)}}}{\partial \hat{\bar{\mu}}_n^{(k)}} \right|_{\hat{\bar{\mu}}_n^{(k)}}, \qquad \mathbb{E}_{f_n^{(k)} q^{\text{old}}}\left[(b_k^{(n)})^2 \right] \approx -2 \left. \frac{\partial \log Z_{f_n^{(k)}}}{\partial \hat{\bar{v}}_n^{(k)}} \right|_{\hat{\bar{v}}_n^{(k)}}, \\
\mathbb{E}_{f_n^{(k)} q^{\text{old}}}\left[1/\sigma_{(k)}^2 \right] &\approx \frac{Z_{f_n^{(k)}}\left(\hat{\bar{\mu}}_n^{(k)}, \hat{\bar{\Sigma}}_n^{(k)}, \hat{\bar{a}}_k + 1, \hat{\bar{b}}_k \right)}{Z_{f_n^{(k)}}\left(\hat{\bar{\mu}}_n^{(k)}, \hat{\bar{\Sigma}}_n^{(k)}, \hat{\bar{a}}_k, \hat{\bar{b}}_k \right)}, \\
\mathbb{E}_{f_n^{(k)} q^{\text{old}}}\left[1/(\sigma_{(k)}^2)^2 \right] &\approx \frac{Z_{f_n^{(k)}}\left(\hat{\bar{\mu}}_n^{(k)}, \hat{\bar{\Sigma}}_n^{(k)}, \hat{\bar{a}}_k + 2, \hat{\bar{b}}_k \right)}{Z_{f_n^{(k)}}\left(\hat{\bar{\mu}}_n^{(k)}, \hat{\bar{\Sigma}}_n^{(k)}, \hat{\bar{a}}_k, \hat{\bar{b}}_k \right)}, \qquad (23)
\end{aligned}
$$

where $Z_{f_n^{(k)}}(\cdot, \cdot, \cdot, \cdot)$ denotes the evaluation of $Z_{f_n^{(k)}}$ with those particular parameters for q^{old}. Note that we are not computing the expected sufficient statistics in the case of the Inverse Gamma distribution over $\sigma_{(k)}^2$, but first and second moments of the inverse variance. These means that we will not minimize the KL-divergence when computing q^{new}. In any case, matching these moments will enforce that the approximate factor $\tilde{f}_n^{(k)}$ is similar to the exact factor in regions of high posterior probability, as indicated by q^{old}. The advantage is that we can compute closed form updates in EP.

Using the moments described above, we can identify the parameters of q^{new} that lead to the

same moments. In particular,

$$\begin{aligned}
\mu_n^{(k)} &= \mathbb{E}_{f_n^{(k)} q^{\text{old}}}\left[b_k^{(n)}\right] , \\
\Sigma_n^{(k)} &= \mathbb{E}_{f_n^{(k)} q^{\text{old}}}\left[(b_k^{(n)})^2\right] - \mathbb{E}_{f_n^{(k)} q^{\text{old}}}\left[b_k^{(n)}\right]^2 , \\
\hat{a}_k &= \frac{\mathbb{E}_{f_n^{(k)} q^{\text{old}}}\left[1/(\sigma_{(k)}^2)^2\right]}{\mathbb{E}_{f_n^{(k)} q^{\text{old}}}\left[1/(\sigma_{(k)}^2)^2\right] - \mathbb{E}_{f_n^{(k)} q^{\text{old}}}\left[1/\sigma_{(k)}^2\right]^2} , \\
\hat{b}_k &= \frac{\hat{a}_k - 1}{\mathbb{E}_{f_n^{(k)} q^{\text{old}}}\left[1/\sigma_{(k)}^2\right]} .
\end{aligned} \tag{24}$$

Given the parameters of q^{new} we can now compute the natural parameters of $\tilde{f}_n^{(n)}$ using the fact that it is proportional to the ratio between q^{new} and q^{old}. Thus, we only have to subtract natural parameters. Namely,

$$\tilde{v}_n^{(k)} = \frac{1}{\Sigma_n^{(k)}} - \hat{\tilde{\Sigma}}_n^{(k)} , \qquad \tilde{m}_n^{(k)} = \frac{\mu_n^{(k)}}{\Sigma_n^{(k)}} - \hat{\tilde{\mu}}_n^{(k)} , \qquad \tilde{a}_n^{(k)} = \hat{a}_k - \hat{\tilde{a}}_k , \qquad \tilde{b}_n^{(k)} = \hat{b}_k - \hat{\tilde{b}}_k . \tag{25}$$

The parameter $\tilde{c}_n^{(k)}$ of $\tilde{f}_n^{(k)}$ is set to guarantee that $f_n^{(k)} q^{\text{old}}$ and $\tilde{f}_n^{(k)} q^{\text{old}}$ integrate the same. Thus, $\tilde{c}_n^{(k)}$ is simply the ratio between the integrals of both terms:

$$\tilde{c}_n^{(k)} = \frac{Z_{f_n^{(k)}} \hat{b}_k^{\hat{a}_k - 1}}{\sqrt{2\pi\Sigma_n^{(k)}} \exp\left\{-\frac{1}{2}\frac{(\mu_n^{(k)})^2}{\Sigma_n^{(k)}}\right\} \Gamma(\hat{a}_k - 1)} . \tag{26}$$

1.3.2 Approximate factors Corresponding to the Robust Prior

Consider the exact factor $g_i^{(k)}(w_i^{(k)}, z_i, \omega_k, \gamma_i, \tau_i^{(k)}, \eta_i^{(k)}) = p(w_i^{(k)}|\mathbf{\Omega}) = \{\pi(w_i^{(k)})^{\eta_i^{(k)}} \delta_0^{1-\eta_i^{(k)}}\}^{z_i} \{[\pi(w_i^{(k)})^{\tau_i^{(k)}} \delta_0^{1-\tau_i^{(k)}}]^{\omega_k} [\pi(w_i^{(k)})^{\gamma_i} \delta_0^{1-\gamma_i}]^{1-\omega_k}\}^{1-z_i}$. To update the corresponding approximate factor $\tilde{g}_i^{(k)}$, consider the the old distribution q^{old}. This distribution is in this case:

$$\begin{aligned}
q^{\text{old}}(w_i^{(k)}, z_i, \omega_k, \gamma_i, \tau_i^{(k)}, \eta_i^{(k)}) = \exp\left\{ \hat{\overline{m}}_i^{(k)} w_i^{(k)} - \frac{\hat{\overline{V}}_{i,i}^{(k)}}{2} (w_i^{(k)})^2 \right\} \exp\{z_i \hat{\overline{p}}_z^{(i)}\} \times \\
\exp\{\omega_k \hat{\overline{p}}_\omega^{(k)}\} \exp\{\gamma_i \hat{\overline{p}}_\gamma^{(i)}\} \exp\{\tau_i^{(k)} \hat{\overline{p}}_\tau^{(i,k)}\} \exp\{\eta_i^{(k)} \hat{\overline{p}}_\eta^{(i,k)}\}
\end{aligned} \tag{27}$$

where $\hat{\overline{m}}_i^{(k)} = m_i^{(k)}/V_{i,i}^{(k)} - \tilde{m}_i^{(k)}$, with $\hat{m}_i^{(k)}$ the i-th component of $\hat{\mathbf{m}}^{(k)}$ and $V_{i,i}^{(k)}$ the i-th diagonal element of $\mathbf{V}^{(k)}$; $\hat{\overline{V}}_{i,i}^{(k)} = 1/V_{i,i}^{(k)} - \tilde{v}_i^{(k)}$; $\hat{\overline{p}}_z^{(i)} = \hat{p}_z^{(i)} - \tilde{p}_z^{(i,k)}$, $\hat{\overline{p}}_\omega^{(k)} = \hat{p}_\omega^{(k)} - \tilde{p}_z^{(i,k)}$, $\hat{\overline{p}}_\gamma^{(i)} = \hat{p}_\gamma^{(i)} - \tilde{p}_\gamma^{(i,k)}$, $\hat{\overline{p}}_\tau^{(i,k)} = \hat{p}_\tau^{(i,k)} - \tilde{p}_\tau^{(i,k)}$ and $\hat{\overline{p}}_\eta^{(i,k)} = \hat{p}_\eta^{(i,k)} - \tilde{p}_\eta^{(i,k)}$. In the previous expression $\tilde{m}_i^{(k)}$, $\tilde{v}_i^{(k)}$, $\tilde{p}_z^{(i,k)}$, $\tilde{p}_\gamma^{(i,k)}$, $\tilde{p}_\tau^{(i,k)}$ and $\tilde{p}_\eta^{(i,k)}$ are the parameters of the approximate factor $\tilde{g}_i^{(k)}$. The other variables are the parameters of q. We are now interested in computing the normalization constant of $g_i^{(k)} q^{\text{old}}$, $Z_{g_i^{(k)}}$. In particular,

$$\begin{aligned}
Z_{g_i^{(k)}} &= \int \sum_{z_i, \omega_k, \gamma_i, \tau_i^{(k)}, \eta_i^{(k)}} g_i^{(k)}(w_i^{(k)}, z_i, \omega_k, \gamma_i, \tau_i^{(k)}, \eta_i^{(k)}) q^{\text{old}}(w_i^{(k)}, z_i, \omega_k, \gamma_i, \tau_i^{(k)}, \eta_i^{(k)}) dw_i^{(k)} \\
&= \exp\{\hat{\overline{p}}_z^{(i)}\} \left\{ \exp\{\hat{\overline{p}}_\eta^{(i,k)}\} Z_\pi + Z_{\delta_0} \right\} + \\
&\quad + \left\{ \exp\{\hat{\overline{p}}_\omega^{(k)}\} \left[\exp\{\hat{\overline{p}}_\eta^{(i,k)}\} Z_\pi + Z_{\delta_0} \right] + \left[\exp\{\hat{\overline{p}}_\gamma^{(i)}\} Z_\pi + Z_{\delta_0} \right] \right\} ,
\end{aligned} \tag{28}$$

where Z_π is given by the convolution of an unnormalized Gaussian distribution and the Strawderman-Bergen prior. This gives

$$
\begin{aligned}
Z_\pi &= \int \pi(w_i^{(k)}) \exp\left\{ \hat{\overline{m}}_i^{(k)} w_i^{(k)} - \frac{\hat{\overline{V}}_{i,i}^{(k)}}{2} (w_i^{(k)})^2 \right\} dw_i^{(k)} \\
&= \frac{1}{\hat{\overline{V}}_{i,i}^{(k)} - 1} \left[\sqrt{\hat{\overline{V}}_{i,i}^{(k)}} \exp\left\{ \frac{1}{2} \frac{\left(\hat{\overline{m}}_i^{(k)}\right)^2}{\hat{\overline{V}}_{i,i}^{(k)}} \right\} - 1 - \right. \\
&\quad \left. - \frac{\sqrt{2\pi} \hat{\overline{m}}_i^{(k)}}{2\sqrt{\hat{\overline{V}}_{i,i}^{(k)} - 1}} \exp\left\{ \frac{1}{2} \frac{\left(\hat{\overline{m}}_i^{(k)}\right)^2}{\hat{\overline{V}}_{i,i}^{(k)} - 1} \right\} \left\{ \operatorname{erf}\left(-C / \sqrt{\hat{\overline{V}}_{i,i}^{(k)}} \right) + \operatorname{erf}(-C) \right\} \right] ,
\end{aligned} \tag{29}
$$

where $C = \hat{\overline{m}}_i^{(k)} / \sqrt{2(\hat{\overline{V}}_{i,i}^{(k)} - 1)}$ and $\operatorname{erf}(\cdot)$ is the error function. This last result extends the one provided in [5], which gives the convolution when $\hat{\overline{V}}_{i,i}^{(k)} = 1$. When $\hat{\overline{V}}_{i,i}^{(k)} < 1$, Z_π is real but its evaluation involves complex numbers. When $\hat{\overline{V}}_{i,i}^{(k)} \to 1$, Z_π tends to the solution given in [5].

Similarly, Z_{δ_0} is the convolution of an unnormalized Gaussian and a delta function centered at zero. That is,

$$
Z_{\delta_0} = \int \delta(w_i^{(k)}) \exp\left\{ \hat{\overline{m}}_i^{(k)} w_i^{(k)} - \frac{\hat{\overline{V}}_{i,i}^{(k)}}{2} (w_i^{(k)})^2 \right\} dw_i^{(k)} = 1 . \tag{30}
$$

Given Z_π and Z_{δ_0}, $Z_{g_i^{(k)}}$ can be readily computed.

The moments of $g_i^{(k)} q^{\text{old}}$ required to find q^{new} are all obtained from $Z_{g_i^{(k)}}$ [6]. In particular, these moments are:

$$
\begin{aligned}
\mathbb{E}_{g_i^{(k)} q^{\text{old}}}\left[w_i^{(k)} \right] &= \left. \frac{\partial \log Z_{g_i^{(k)}}}{\partial \hat{\overline{m}}_i^{(k)}} \right|_{\hat{\overline{m}}_i^{(k)}} , & \mathbb{E}_{g_i^{(k)} q^{\text{old}}}\left[(w_i^{(k)})^2 \right] &= -2 \left. \frac{\partial \log Z_{g_i^{(k)}}}{\partial \hat{\overline{v}}_i^{(k)}} \right|_{\hat{\overline{v}}_i^{(k)}} , \\
\mathbb{E}_{g_i^{(k)} q^{\text{old}}}\left[z_i \right] &= \left. \frac{\partial \log Z_{g_i^{(k)}}}{\partial \hat{\overline{p}}_z^{(i)}} \right|_{\hat{\overline{p}}_z^{(i)}} , & \mathbb{E}_{g_i^{(k)} q^{\text{old}}}\left[\omega_k \right] &= \left. \frac{\partial \log Z_{g_i^{(k)}}}{\partial \hat{\overline{p}}_\omega^{(k)}} \right|_{\hat{\overline{p}}_\omega^{(k)}} , \\
\mathbb{E}_{g_i^{(k)} q^{\text{old}}}\left[\gamma_i \right] &= \left. \frac{\partial \log Z_{g_i^{(k)}}}{\partial \hat{\overline{p}}_\gamma^{(i)}} \right|_{\hat{\overline{p}}_\gamma^{(i)}} , & \mathbb{E}_{g_i^{(k)} q^{\text{old}}}\left[\tau_i^{(k)} \right] &= \left. \frac{\partial \log Z_{g_i^{(k)}}}{\partial \hat{\overline{p}}_\tau^{(i,k)}} \right|_{\hat{\overline{p}}_\tau^{(i,k)}} , \\
\mathbb{E}_{g_i^{(k)} q^{\text{old}}}\left[\eta_i^{(k)} \right] &= \left. \frac{\partial \log Z_{g_i^{(k)}}}{\partial \hat{\overline{p}}_\eta^{(i,k)}} \right|_{\hat{\overline{p}}_\eta^{(i,k)}} .
\end{aligned} \tag{31}
$$

Using the moments described above, we can identify the parameters of q^{new} that lead to the same moments. In particular,

$$
\begin{aligned}
m_i^{(k)} &= \mathbb{E}_{g_i^{(k)} q^{\text{old}}}\left[w_i^{(k)} \right] , & V_{i,i}^{(k)} &= \mathbb{E}_{g_i^{(k)} q^{\text{old}}}\left[(w_i^{(k)})^2 \right] - \mathbb{E}_{g_i^{(k)} q^{\text{old}}}\left[w_i^{(k)} \right]^2 , \\
p_z^{(i)} &= \mathbb{E}_{g_i^{(k)} q^{\text{old}}}\left[z_i \right] , & p_\omega^{(k)} &= \mathbb{E}_{g_i^{(k)} q^{\text{old}}}\left[\omega_k \right] , \\
p_\gamma^{(i)} &= \mathbb{E}_{g_i^{(k)} q^{\text{old}}}\left[\gamma_i \right] , & p_\tau^{(i,k)} &= \mathbb{E}_{g_i^{(k)} q^{\text{old}}}\left[\tau_i^{(k)} \right] , \\
p_\eta^{(i,k)} &= \mathbb{E}_{g_i^{(k)} q^{\text{old}}}\left[\eta_i^{(k)} \right] .
\end{aligned} \tag{32}
$$

Given the parameters of q^{new} we can now compute the natural parameters of $\tilde{g}_i^{(k)}$ using the fact that it is proportional to the ratio between q^{new} and q^{old}. Thus, we only have to subtract natural parameters. Namely,

$$\begin{aligned}
\tilde{m}_i^{(k)} &= \frac{m_i^{(k)}}{V_{i,i}^{(k)}} - \hat{\overline{m}}_i^{(k)}\,, & \tilde{v}_i^{(k)} &= \frac{1}{V_{i,i}^{(k)}} - \hat{\overline{V}}_{i,i}^{(k)}\,, \\
\tilde{p}_z^{(i,k)} &= \log\left(\frac{p_z^{(i)}}{1-p_z^{(i)}}\right) - \hat{\overline{p}}_z^{(i,k)}\,, & \tilde{p}_\omega^{(i,k)} &= \log\left(\frac{p_\omega^{(k)}}{1-p_\omega^{(k)}}\right) - \hat{\overline{p}}_\omega^{(i,k)}\,, \\
\tilde{p}_\gamma^{(i,k)} &= \log\left(\frac{p_\gamma^{(i)}}{1-p_\gamma^{(k)}}\right) - \hat{\overline{p}}_\gamma^{(i,k)}\,, & \tilde{p}_\tau^{(i,k)} &= \log\left(\frac{p_\tau^{(i,k)}}{1-p_\tau^{(i,k)}}\right) - \hat{\overline{p}}_\tau^{(i,k)}\,, \\
\tilde{p}_\eta^{(i,k)} &= \log\left(\frac{p_\eta^{(i,k)}}{1-p_\eta^{(i,k)}}\right) - \hat{\overline{p}}_\eta^{(i,k)}\,.
\end{aligned} \tag{33}$$

The parameter $\tilde{s}_i^{(k)}$ of $\tilde{g}_i^{(k)}$ is set to guarantee that $g_i^{(k)} q^{\text{old}}$ and $\tilde{g}_i^{(k)} q^{\text{old}}$ integrate the same. Thus, $\tilde{s}_i^{(k)}$ is simply the ratio between the integrals of both terms:

$$\begin{aligned}
\tilde{s}_i^{(k)} = \frac{Z_{\tilde{g}_i^{(k)}}}{\sqrt{2\pi V_{i,i}^{(k)}}\exp\left\{-\frac{1}{2}\frac{(m_i^{(k)})^2}{V_{i,i}^{(k)}}\right\}} \cdot \frac{1}{(1+\exp(\tilde{p}_z^{(i,k)}))} \cdot \frac{1}{(1+\exp(\tilde{p}_\omega^{(i,k)}))} \\
\frac{1}{(1+\exp(\tilde{p}_\gamma^{(i,k)}))} \cdot \frac{1}{(1+\exp(\tilde{p}_\tau^{(i,k)}))} \cdot \frac{1}{(1+\exp(\tilde{p}_\eta^{(i,k)}))}\,.
\end{aligned} \tag{34}$$

where we have used the natural and standard parameters of q^{new} that result from computing the product $\tilde{g}_i^{(k)}$.

1.3.3 Approximate Factors Corresponding to the Beta Priors

In this section we describe how to process the factor $p(z_i|\rho_z) = h_z^{(i)}(z_i, \rho_z) = \text{Bernoulli}(z_i|\rho_z)$. Processing the factors corresponding to the priors of the other binary latent variables is very similar and hence omitted. To update the corresponding approximate factor $\tilde{h}_z^{(i)}$, consider the corresponding old distribution q^{old}. This distribution is:

$$q^{\text{old}}(z_i, \rho_z) = \exp(z_i \hat{\overline{p}}_z^{(i)}) \exp(\hat{\overline{a}}_z^{(i)} \log(\rho_z) + \hat{\overline{b}}_z^{(i)} \log(1-\rho_z))\,, \tag{35}$$

where $\hat{\overline{p}}_z^{(i)} = \hat{p}_z^{(i)} - \tilde{p}_z^{(i)}$, $\hat{\overline{a}}_z^{(i)} = \hat{a}_z^{(i)} - \tilde{a}_z^{(i)}$, and $\hat{\overline{b}}_z^{(i)} = \hat{b}_z^{(i)} - \tilde{b}_z^{(i)}$. In the previous expressions $\tilde{p}_z^{(i)}$, $\tilde{a}_z^{(i)}$ and $\tilde{b}_z^{(i)}$ are the parameters of $\tilde{h}_z^{(i)}$. The other variables are the natural parameters of q. We compute now the normalization constant of $h_z^{(i)} q^{\text{old}}$, $Z_{h_z^{(i)}}$. In particular,

$$\begin{aligned}
Z_{h_z^{(i)}} &= \int \sum_{z_i} q^{\text{old}}(z_i, \rho_z) \text{Bernoulli}(z_i|\rho_z) d\rho_z \\
&= \int \sum_{z_i} q^{\text{old}}(z_i, \rho_z) \rho_z^{z_i} (1-\rho_z)^{1-z_i} d\rho_z \\
&= \exp(z_i \hat{\overline{p}}_z^{(i)}) \int \exp((\hat{\overline{a}}_z^{(i)} + 1)\log(\rho_z) + \hat{\overline{b}}_z^{(i)} \log(1-\rho_z)) d\rho_z + \\
&\quad + \int \exp(\hat{\overline{a}}_z^{(i)} \log(\rho_z) + (\hat{\overline{b}}_z^{(i)} + 1)\log(1-\rho_z)) d\rho_z \\
&= \exp(z_i \hat{\overline{p}}_z^{(i)}) \beta(\hat{\overline{a}}_z^{(i)} + 2, \hat{\overline{b}}_z^{(i)} + 1) + \beta(\hat{\overline{a}}_z^{(i)} + 1, \hat{\overline{b}}_z^{(i)} + 2)\,,
\end{aligned} \tag{36}$$

where $\beta(\cdot,\cdot)$ is the beta function.

The moments of $h_z^{(i)} q^{\text{old}}$ required to find q^{new} are all obtained from $Z_{h_z^{(i)}}$ [6]. In particular, these moments are:

$$\mathbb{E}_{h_z^{(i)} q^{\text{old}}}\left[\rho_z\right] = \frac{Z_{h_z^{(i)}}(\hat{\overline{p}}_z^{(i)}, \hat{\overline{a}}_z^{(i)} + 1, \hat{\overline{b}}_z^{(i)})}{Z_{h_z^{(i)}}}, \qquad \mathbb{E}_{h_z^{(i)} q^{\text{old}}}\left[\rho_z^2\right] = \frac{Z_{h_z^{(i)}}(\hat{\overline{p}}_z^{(i)}, \hat{\overline{a}}_z^{(i)} + 2, \hat{\overline{b}}_z^{(i)})}{Z_{h_z^{(i)}}},$$
$$\mathbb{E}_{h_z^{(i)} q^{\text{old}}}\left[z_i\right] = \left.\frac{\partial \log Z_{h_z^{(i)}}}{\partial \hat{\overline{p}}_z^{(i)}}\right|_{\hat{\overline{p}}_z^{(i)}}. \tag{37}$$

where $Z_{h_z^{(i)}}(\cdot,\cdot,\cdot)$ indicates the evaluation of $Z_{h_z^{(i)}}$ with the natural parameters for q^{old} specified in its arguments. Note again that we are not computing the expected sufficient statistics in the case of the Beta distribution over ρ_z, but first and second moments. These means that we will not minimize the KL-divergence when computing q^{new}. In any case, matching these moments will enforce that the approximate factor $\tilde{h}_z^{(i)}$ is similar to the exact factor in regions of high posterior probability, as indicated by q^{old}. This has also been done, for example, in [1]. The advantage is that we can compute closed form updates in EP.

Using the moments described above, we can identify the parameters of q^{new} that lead to the same moments. In particular,

$$\begin{aligned} p_z^{(i)} &= \mathbb{E}_{h_z^{(i)} q^{\text{old}}}\left[z_i\right], \\ a_z &= \frac{\mathbb{E}_{h_z^{(i)} q^{\text{old}}}\left[\rho_z\right]\left(\mathbb{E}_{h_z^{(i)} q^{\text{old}}}\left[\rho_z\right] - \mathbb{E}_{h_z^{(i)} q^{\text{old}}}\left[\rho_z^2\right]\right)}{\mathbb{E}_{h_z^{(i)} q^{\text{old}}}\left[\rho_z^2\right] - \mathbb{E}_{h_z^{(i)} q^{\text{old}}}\left[\rho_z\right]^2}, \\ b_z &= \frac{\left(1 - \mathbb{E}_{h_z^{(i)} q^{\text{old}}}\left[\rho_z\right]\right)\left(\mathbb{E}_{h_z^{(i)} q^{\text{old}}}\left[\rho_z\right] - \mathbb{E}_{h_z^{(i)} q^{\text{old}}}\left[\rho_z^2\right]\right)}{\mathbb{E}_{h_z^{(i)} q^{\text{old}}}\left[\rho_z^2\right] - \mathbb{E}_{h_z^{(i)} q^{\text{old}}}\left[\rho_z\right]^2}. \end{aligned} \tag{38}$$

Given the parameters of q^{new} we can now compute the natural parameters of $\tilde{h}_z^{(i)}$ using the fact that it is proportional to the ratio between q^{new} and q^{old}. Thus, we only have to subtract natural parameters. Namely,

$$\tilde{p}_z^{(i)} = \log\left(\frac{p_z^{(i)}}{1 - p_z^{(i)}}\right) - \hat{\overline{p}}_z^{(i)}, \qquad \tilde{a}_z^{(i)} = a_z - 1 - \hat{\overline{a}}_z^{(i)}, \qquad \tilde{b}_z^{(i)} = b_z - 1 - \hat{\overline{b}}_z^{(i)}. \tag{39}$$

The parameter $\tilde{k}_z^{(i)}$ of $\tilde{h}_z^{(i)}$ is set to guarantee that $h_z^{(i)} q^{\text{old}}$ and $\tilde{h}_z^{(i)} q^{\text{old}}$ integrate the same. Thus, $\tilde{k}_z^{(i)}$ is simply the ratio between the integrals of both terms:

$$\tilde{k}_z^{(i)} = \frac{Z_{h_z^{(i)}}}{\beta(a_z, b_z)\left(1 + \exp\left\{\hat{p}_z^{(i)}\right\}\right)} \tag{40}$$

where we have used the natural and standard parameters of q^{new} that result from computing the product $\tilde{h}_z^{(i)} q^{\text{old}}$.

1.4 Approximation of the Marginal Likelihood

The marginal likelihood can be approximated by the normalization constant of the approximate joint distribution $\tilde{q}$. That is,

$$
\begin{aligned}
\text{ML} &\approx \int \sum_{\boldsymbol{\Omega}} \tilde{q}(\mathbf{W}, \boldsymbol{\Omega}, \boldsymbol{\rho}, \boldsymbol{\sigma}^2) d\mathbf{W} d\boldsymbol{\sigma}^2 d\boldsymbol{\rho} \\
&= \left[\prod_{n=1}^{N_k} \prod_{k=1}^{K} \tilde{c}_n^{(k)} \right] \left[\prod_{i=1}^{d} \prod_{k=1}^{K} \tilde{s}_i^{(k)} \right] \left[\prod_{i=1}^{d} \tilde{k}_z^{(i)} \right] \left[\prod_{k=1}^{K} \tilde{k}_\omega^{(k)} \right] \left[\prod_{i=1}^{d} \tilde{k}_\gamma^{(i)} \right] \times \\
&\quad \times \left[\prod_{k=1}^{K} \prod_{i=1}^{d} \tilde{k}_\tau^{(i,k)} \right] \left[\prod_{k=1}^{K} \prod_{i=1}^{d} \tilde{k}_\eta^{(i,k)} \right] \left[\prod_{k=1}^{K} (2\pi)^{\frac{d}{2}} \sqrt{|\mathbf{V}^{(k)}|} \exp\left\{ \frac{1}{2} (\mathbf{m}^{(k)})^{\mathrm{T}} (\mathbf{V}^{(k)})^{-1} \mathbf{m}^{(k)} \right\} \right] \times \\
&\quad \times \left[\prod_{i=1}^{d} \left(1 + \exp\left\{ \hat{p}_z^{(i)} \right\} \right) \right] \left[\prod_{i=1}^{d} \left(1 + \exp\left\{ \hat{p}_\gamma^{(i)} \right\} \right) \right] \left[\prod_{k=1}^{K} \left(1 + \exp\left\{ \hat{p}_\omega^{(i)} \right\} \right) \right] \times \\
&\quad \times \left[\prod_{k=1}^{K} \prod_{i=1}^{d} \left(1 + \exp\left\{ \hat{p}_\tau^{(i,k)} \right\} \right) \right] \left[\prod_{k=1}^{K} \prod_{i=1}^{d} \left(1 + \exp\left\{ \hat{p}_\eta^{(i,k)} \right\} \right) \right] \times \\
&\quad \times \frac{\beta(a_z, b_z) \beta(a_\omega, b_\omega) \beta(a_\gamma, b_\gamma) \beta(a_\tau, b_\tau) \beta(a_\eta, b_\eta)}{\beta(1,1)^5} \prod_{k=1}^{K} \frac{\Gamma(a_k) 5^5}{\Gamma(5) {b_k}^{a_k}} \,, \qquad (41)
\end{aligned}
$$

where we have used the fact that $\tilde{q}$ is the posterior approximation q without normalization. Thus, the approximation to the marginal likelihood is expressed in terms of the parameters (natural and standard) of q. Furthermore, all operations involving $\mathbf{V}^{(k)}$ can be carried out with cost $\mathcal{O}(N_k^2 d)$, using the special structure of this matrix. For example, $|\mathbf{V}^{(k)}|$ can be efficiently computed using Sylvester's determinant theorem. The total cost of computing the approximation to the marginal likelihood is $\mathcal{O}(\sum_{k=1}^{K} N_k^2 d)$.

2 Denoising of Natural Images

In this section we show a representative example of the 64 different groups of non-overlapping blocks considered in the experiments involving the denoising of the house image. Recall that there is 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. Figure 1 shows a representative example of each these groups. The group that contains 32×32 blocks, a group that contains 32×31 blocks, a group that contains 31×32 blocks and, finally, a group that contains 31×31 blocks.

Recall that each group of non-overlapping blocks is regarded as a multi-task learning problem with as many tasks as blocks. In particular, $\mathbf{y}^{(k)} = \mathbf{X}^{(k)} \mathbf{w}^{(k)} + \boldsymbol{\epsilon}^{(k)}$, where $\mathbf{y}^{(k)}$ denotes a particular block, $\mathbf{X}^{(k)}$ is a wavelet basis corresponding to the Haar wavelet and $\boldsymbol{\epsilon}^{(k)}$ is additive Gaussian noise. The main advantage of using these groups in the learning process is that each multi-task learning problem corresponding to each of the 64 groups of non-overlapping blocks can be solved in parallel using each multi-task method. This is specially convenient because some of the methods we compare with are particularly slow, *e.g.*, MTFS_{Dep}.

Given an estimate of $\mathbf{w}^{(k)}$, $\hat{\mathbf{w}}^{(k)}$, these coefficients can be projected onto $\mathbf{X}^{(k)}$, *i.e.*, by computing $\mathbf{X}^{(k)} \hat{\mathbf{w}}^{(k)}$, to get an estimate of the original noise-less block. After doing this for each block, the original image can be reconstructed simply by carefully averaging the different blocks as in [7].

References

[1] D. Hernández-Lobato and J. M. Hernández-Lobato. Bayes machines for binary classification. *Pattern Recognition Letters*, 29(10):1466–1473, 2008.

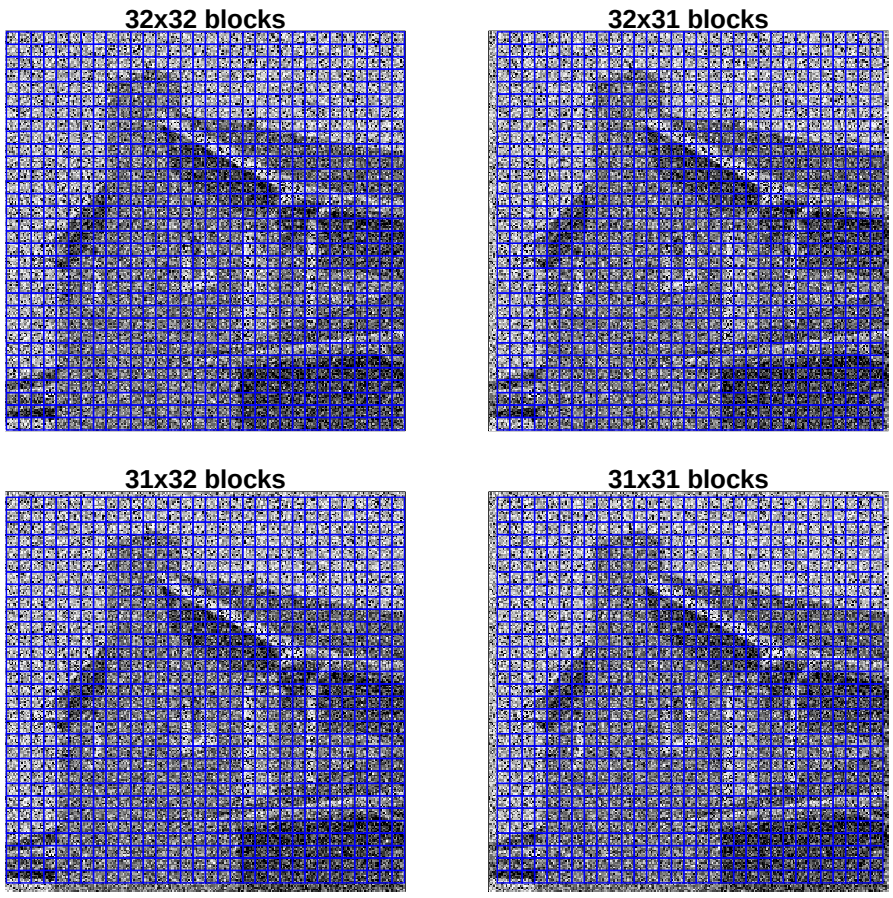

Figure 1: This figure shows a representative subset of the 64 different groups of non-overlapping blocks considered in the experiments involving the denoising of the house image.

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

[3] J. M. Hernndez-Lobato, N. Houlsby, and Z. Ghahramani. Probabilistic matrix factorization with non-random missing data. In *International Conference on Machine Learning*, pages 1512–1520, 2014.

[4] N. Houlsby, J. M. Hernndez-Lobato, and Z. Ghahramani. Cold-start active learning with robust ordinal matrix factorization. In *International Conference on Machine Learning*, pages 766–774, 2014.

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

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

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

[8] M. Van Gerven, B. Cseke, Oostenveld R., and T. Heskes. Bayesian source localization with the multivariate Laplace prior. In *Neural Information Processing Systems*, pages 1901–1909, 2009.