Journal of Machine Learning Research 14 (2013) 1891-1945 Submitted 10/12; Revised 2/13; Published 7/13

Generalized Spike-and-Slab Priors for Bayesian Group Feature Selection Using Expectation Propagation

Daniel Hernández-Lobato DANIEL.HERNANDEZ@UAM.ES
Computer Science Department
Universidad Autónoma de Madrid
Calle Francisco Tomás y Valiente 11
28049 Madrid, Spain

José Miguel Hernández-Lobato JMH233@ENG.CAM.AC.UK
Department of Engineering
University of Cambridge
Trumpington Street
Cambridge CB2 1PZ, United Kingdom

Pierre Dupont PIERRE.DUPONT@UCLOUVAIN.BE
Machine Learning Group, ICTEAM
Université catholique de Louvain
Place Sainte Barbe 2
1348, Louvain-la-Neuve, Belgium

Editor: David Blei

Abstract

We describe a Bayesian method for group feature selection in linear regression problems. The method is based on a generalized version of the standard spike-and-slab prior distribution which is often used for individual feature selection. Exact Bayesian inference under the prior considered is infeasible for typical regression problems. However, approximate inference can be carried out efficiently using Expectation Propagation (EP). A detailed analysis of the generalized spike-and-slab prior shows that it is well suited for regression problems that are sparse at the group level. Furthermore, this prior can be used to introduce prior knowledge about specific groups of features that are *a priori* believed to be more relevant. An experimental evaluation compares the performance of the proposed method with those of group LASSO, Bayesian group LASSO, automatic relevance determination and additional variants used for group feature selection. The results of these experiments show that a model based on the generalized spike-and-slab prior and the EP algorithm has state-of-the-art prediction performance in the problems analyzed. Furthermore, this model is also very useful to carry out sequential experimental design (also known as active learning), where the data instances that are most informative are iteratively included in the training set, reducing the number of instances needed to obtain a particular level of prediction accuracy.

Keywords: group feature selection, generalized spike-and-slab priors, expectation propagation, sparse linear model, approximate inference, sequential experimental design, signal reconstruction

1. Introduction

Many regression problems of practical interest are characterized by a small number of training instances n and a large number of explanatory variables or features d. Examples of these problems

©2013 Daniel Hernández-Lobato, José Miguel Hernández-Lobato and Pierre Dupont.

include the reconstruction of medical images (Seeger et al., 2009), the processing of fMRI data (Gerven et al., 2009), the discovery of gene regulators (Hernández-Lobato et al., 2008), the processing of natural language (Sandler et al., 2009) or the reconstruction of transcription networks (Hernández-Lobato and Dijkstra, 2010). Under these circumstances a simple linear model is typically assumed to describe the observed data. However, when $n \ll d$ the problem of estimating the optimal linear relation is under-determined. In particular, there is an infinite number of values for the model coefficients that explain the data equally well (Johnstone and Titterington, 2009). To address this difficulty and also to alleviate over-fitting, a practical solution that is typically employed is to assume that only a few of all the explanatory variables or features are actually relevant for prediction. The consequence is that the estimation problem is generally regularized by assuming that most of the model coefficients in the optimal solution are exactly equal to zero (Johnstone and Titterington, 2009). That is, the vector of model coefficients is sparse. This means that the features whose associated model coefficients take value equal to zero do not contribute to the decisions made by the model and are hence considered to be irrelevant. To estimate the model coefficients under the sparsity assumption different strategies described in the literature can be used. These include introducing regularization norms or assuming sparse enforcing priors in the estimation process (Tibshirani, 1996; George and McCulloch, 1997; Tipping, 2001; Kappen and Gómez, 2013).

The process of inducing the model coefficients under the sparsity assumption can be facilitated when prior information is available about groups of features that are expected to be jointly relevant or jointly irrelevant for prediction (Huang and Zhang, 2010). That is, when different groups of model coefficients are expected to be jointly equal to or jointly different from zero. Finding this type of information can be difficult in practice. However, such prior information can be deduced from a related task (see, e.g., Section 7.2) or from some additional data (see, e.g., Section 7.3). Furthermore, if this information is available and it is accurate, it can be beneficial to improve the estimates of the model coefficients and to reduce the number of samples required to obtain a good generalization performance. As described by Puig et al. (2011) the class of problems where sparsity at the group level is beneficial include spectrum cartography for cognitive radio (Bazerque et al., 2011), jointly-sparse signal recovery (Wakin et al., 2006), regression with grouped variables (Yuan and Lin, 2006) and source localization (Malioutov et al., 2005). Other classes of problems that can benefit from group sparsity are multi-task feature selection (Hernández-Lobato et al., 2010) or whole genome association mapping (Kim and Xing, 2008). As with the individual sparsity assumption, sparsity at the group level can be introduced in the estimation process of the model coefficients by considering specific regularization norms or by assuming sparse enforcing priors at the group level (Yuan and Lin, 2006; Ji et al., 2009; Vogt and Roth, 2010; Raman et al., 2009; Yen and Yen, 2011). For this purpose, we specifically consider a generalized version of the standard spike-and-slab prior which has been typically employed for individual feature selection (Mitchell and Beauchamp, 1988; Geweke, 1996; George and McCulloch, 1997). Under the assumption that the grouping information is given beforehand, the proposed prior introduces a set of binary latent variables, one for each different group of features. If the latent variable of a particular group is equal to zero, the model coefficients corresponding to that group are set to zero and the features of the group are not used for prediction of the targets. On the other hand, if the latent variable is equal to one, the features of that particular group are used for prediction and the model coefficients are assumed to be generated from a multi-variate Gaussian distribution. When there is no grouping information this prior reduces to the standard spike-and-slab prior. Exact Bayesian inference under the prior considered is infeasible

for typical regression problems. Thus, in practice one has to resort to approximate techniques for Bayesian inference.

The proposed generalized version of the spike-and-slab prior has several practical advantages over other methods for group feature selection. In particular, it is the only prior that puts a positive probability mass on values equal to zero for the model coefficients of each group. Furthermore, introducing the prior fraction of relevant groups for prediction or the expected deviation from zero of the coefficients that are actually different from zero is very easy under this prior. The proposed prior also has a closed form convolution with the Gaussian distribution, which allows the use of efficient algorithms for approximate Bayesian inference like expectation propagation (EP) (Minka, 2001). This is impossible with other existing priors that can also be used for group feature selection like the multi-variate Laplace distribution (Raman et al., 2009) or the multi-variate horseshoe (Carvalho et al., 2009). The proposed prior also provides a direct estimate of the relative importance of each group for prediction, hence identifying easily the most relevant groups. This estimate is simply obtained by computing the posterior probability that the different latent variables of the prior are activated.

The EP algorithm has a total running time under this prior that scales as $O(n^2 d)$, where n is the number of training instances and d is the number of features. This linear dependence on d is very efficient when $n \ll d$, which is the general scenario we assume. The EP algorithm also provides a direct estimate of the posterior covariances of the model coefficients. These are useful for carrying out sequential experimental design in a linear model that incorporates the grouping information using the proposed generalized spike-and-slab prior. Specifically, in sequential experimental design we save on costly experiments by iteratively including in the training set the data instances that are most informative about the model coefficients (Seeger, 2008). Our experiments indicate that EP and the proposed prior are very effective towards this end. Additionally, a detailed analysis of the sparsity properties of the generalized prior shows that it is adequate for group feature selection. In particular, it produces selective shrinkage of the different groups of the model coefficients. Namely, under this prior it is possible to achieve high levels of group sparsity and, at the same time, to avoid shrinking the model coefficients that are truly different from zero. We show that this is not possible with other popular methods for group feature selection. Finally, incorporating prior knowledge about specific groups of features that are more likely to be used for prediction is straight-forward under the generalized spike-and-slab prior. This type of prior information can be very useful to improve the prediction performance.

The performance of a model based on the generalized spike-and-slab prior and the EP algorithm is evaluated on a collection of benchmark regression problems and compared with other methods from the literature that can also be used for group feature selection. The problems investigated include the reconstruction of sparse signals from a reduced set of noisy measurements (Ji et al., 2008), the prediction of user sentiment (Blitzer et al., 2007), and the reconstruction of images of hand-written digits extracted from the MNIST data set (LeCun et al., 1998). The methods we compare with include the group LASSO (Yuan and Lin, 2006; Kim et al., 2006), the Bayesian group LASSO (Raman et al., 2009), a modified version of the horseshoe prior for group feature selection (Carvalho et al., 2009) and the group automatic relevance determination (ARD) principle (Ji et al., 2009). We also include for comparison a model which does not consider the grouping information and a model that uses Markov chain Monte Carlo methods (Gibbs sampling) for approximate inference instead of the EP algorithm. The results of these experiments indicate that the grouping information can significantly improve the performance of the prediction models. Furthermore, the

model based on the proposed prior and the EP algorithm generally performs best in the problems investigated. There is little difference between the performance obtained by using EP or Gibbs sampling for approximate inference. This confirms the accuracy of the posterior approximation computed by EP, while the EP algorithm is orders of magnitude faster than Gibbs sampling. The running time of EP is also better than or similar to the running time of the other methods analyzed in this document. These experiments also show that the proposed prior is very useful to identify the most relevant groups for prediction and that prior knowledge about specific groups that are *a priori* expected to be more relevant for prediction can significantly improve the prediction performance of the resulting model.

The rest of this document is organized as follows. Section 2 describes the generalized spike-and-slab prior considered for group feature selection. Section 3 shows how the EP algorithm can be successfully applied for approximate inference in a linear regression model based on this prior. Section 4 introduces sequential experimental design and shows how it can be efficiently implemented in the model based on the generalized spike-and-slab prior and the EP algorithm. Section 5 gives a summary of other methods that are available in the literature to carry out group feature selection. Section 6 introduces a detailed analysis of the group sparsity properties of the generalized spike-and-slab prior and the other methods that can be used for this purpose. Section 7 shows some experiments comparing the different methods that can be used for group feature selection and, finally, Section 8 presents the conclusions of this work.

2. Group Feature Selection Using Spike-and-Slab Priors

In this section we describe the linear regression model which promotes sparsity at the group level using a generalized spike-and-slab prior. Consider some training data in the form of n d-dimensional feature vectors summarized in a design matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathrm{T}}$ and some associated target values $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ with $y_i \in \mathbb{R}$. A linear predictive rule is assumed for $\mathbf{y}$ given $\mathbf{X}$:

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}\,, \tag{1}$$

where $\mathbf{w}$ is a vector of unknown model coefficients and $\boldsymbol{\epsilon}$ is a n-dimensional vector of independent additive Gaussian noise with variance σ_0^2, that is, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_0^2\mathbf{I})$. Given $\mathbf{X}$ and $\mathbf{y}$, the likelihood of $\mathbf{w}$ is defined as

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^{n} p(y_i|\mathbf{x}_i, \mathbf{w}) = \prod_{i=1}^{n} \mathcal{N}(y_i|\mathbf{w}^{\mathrm{T}}\mathbf{x}_i, \sigma_0^2)\,. \tag{2}$$

When $d \gg n$, (2) is not strictly concave and infinitely many values of $\mathbf{w}$ fit the training data equally well with perfect prediction accuracy. These are precisely the type of problems we are interested in. A strong regularization technique that is typically employed in this context is to assume that $\mathbf{w}$ is sparse. We further assume the availability of prior information about groups of components of $\mathbf{w}$ that are expected to be jointly zero or jointly different from zero. This is equivalent to considering specific groups of features that are expected to be jointly relevant or jointly irrelevant for prediction. All these assumptions can be incorporated into the model using a generalized version of the standard *spike-and-slab* prior (Mitchell and Beauchamp, 1988; Geweke, 1996; George and McCulloch, 1997).

Consider a partition of $\mathbf{w}$ into G disjoint groups (in general we do not allow for groups of model coefficients to overlap) such that $\mathbf{w} = (\mathbf{w}_1^{\mathrm{T}}, \ldots, \mathbf{w}_G^{\mathrm{T}})^{\mathrm{T}}$. We introduce a vector $\mathbf{z}$ with G binary latent

variables $\{z_1,\ldots,z_G\}$, where each z_g indicates whether $\mathbf{w}_g$ is zero ($z_g = 0$) or different from zero ($z_g = 1$). When $\mathbf{z}$ is known, the prior for $\mathbf{w}$ is defined as:

$$\begin{aligned}\mathcal{P}(\mathbf{w}|\mathbf{z}) &= \prod_{g=1}^{G} \left[z_g \mathcal{N}(\mathbf{w}_g|\mathbf{0}, v_0\mathbf{I}) + (1-z_g)\delta(\mathbf{w}_g)\right] \\ &= \prod_{j=1}^{d} \left[z_{g(j)} \mathcal{N}(w_j|0, v_0) + (1-z_{g(j)})\delta(w_j)\right], \end{aligned} \tag{3}$$

where $g(j)$ is the index of the group which contains the j-th coefficient, $\mathcal{N}(\cdot|\mathbf{0}, v_0\mathbf{I})$ is a Gaussian density, with zero mean and a group specific variance v_0 (the slab), and $\delta(\cdot)$ is a point probability mass centered at the origin (the spike). The value of v_0 controls the shrinkage of the coefficients that are different from zero. If v_0 is large, the coefficients of the groups that are different from zero are barely regularized. Conversely, if v_0 is small these coefficients are strongly shrunk towards zero. Finally, the prior for $\mathbf{z}$ is a multivariate Bernoulli distribution:

$$\mathcal{P}(\mathbf{z}) = \text{Bern}(\mathbf{z}|\mathbf{p}_0) = \prod_{g=1}^{G} \left[p_{0,g}^{z_g}(1-p_{0,g})^{(1-z_g)}\right], \tag{4}$$

where $p_{0,g}$ is the prior probability that the coefficients within the g-th group are different from zero and $\mathbf{p}_0 = (p_{0,1},\ldots,p_{0,G})^{\mathrm{T}}$. Thus, incorporating prior knowledge about specific groups of features that are more likely to be used for prediction is straight-forward under (4). For this we only have to increase the corresponding components of the vector $\mathbf{p}_0$. When all groups of features are *a priori* believed to be equally relevant for prediction, each $p_{0,g}$ with $g = 1,\ldots,G$ can be set equal to a constant value p_0, which indicates the fraction of groups initially expected to be relevant for prediction. Finally, we note that when all groups of features are of size one, (3) reduces to the standard spike-and-slab prior described by George and McCulloch (1997).

2.1 Inference, Prediction and Relevant Groups

Given the observed data $\mathbf{X}$ and $\mathbf{y}$, we make inference about the potential values of $\mathbf{w}$ and $\mathbf{z}$ using Bayes' theorem:

$$\mathcal{P}(\mathbf{w},\mathbf{z}|\mathbf{y},\mathbf{X}) = \frac{\mathcal{P}(\mathbf{y}|\mathbf{w},\mathbf{X})\mathcal{P}(\mathbf{w}|\mathbf{z})\mathcal{P}(\mathbf{z})}{\mathcal{P}(\mathbf{y}|\mathbf{X})}, \tag{5}$$

where $\mathcal{P}(\mathbf{y}|\mathbf{X})$ is a normalization constant, known as the model evidence, which is useful for model comparison (Bishop, 2006; MacKay, 2003). This posterior distribution and the likelihood (2) can be combined to compute a predictive distribution for the target $y_{\text{new}} \in \mathbb{R}$ associated to a new observation $\mathbf{x}_{\text{new}}$:

$$\mathcal{P}(y_{\text{new}}|\mathbf{x}_{\text{new}},\mathbf{y},\mathbf{X}) = \sum_{\mathbf{z}} \int \mathcal{P}(y_{\text{new}}|\mathbf{w},\mathbf{x}_{\text{new}})\mathcal{P}(\mathbf{w},\mathbf{z}|\mathbf{y},\mathbf{X})\, d\mathbf{w}. \tag{6}$$

The posterior distribution of $\mathbf{z}$ defines the probability of the features contained in specific relevant groups to be used for prediction:

$$\mathcal{P}(\mathbf{z}|\mathbf{y},\mathbf{X}) = \int \mathcal{P}(\mathbf{w},\mathbf{z}|\mathbf{y},\mathbf{X})\, d\mathbf{w}. \tag{7}$$

Furthermore, one can marginalize (7) over $z_{g'}$, with $g' \neq g$, for a specific latent variable z_g to compute $\mathcal{P}(z_g|\mathbf{y},\mathbf{X})$, that is, the associated posterior probability of using the g-th group of features for prediction.

A practical difficulty is that the exact computation of (5), (6) and (7) is intractable for typical learning problems. All these expressions involve a number of summations which grows exponentially with G, which is typically of the same order as d. Thus, they must be approximated in practice. Approximate Bayesian inference is typically implemented in the literature using Markov chain Monte Carlo techniques (Gibbs sampling, in particular) where one samples from a Markov chain whose stationary distribution coincides with the posterior distribution of the model (Neal, 1993). Unfortunately, these methods are computationally expensive since the Markov chain has to be run for a large number of iterations to get just a few independent samples. As a more efficient alternative we employ here expectation propagation (EP), a method for fast approximate inference with Bayesian models (Minka, 2001). This method is described in the next section.

3. Expectation Propagation for Bayesian Group Feature Selection

Expectation propagation (EP) is a deterministic method for carrying out approximate Bayesian inference (Minka, 2001). EP approximates the posterior distribution of the parameters of interest using a simpler parametric distribution Q. The form of Q is chosen so that the integrals required to calculate expected values and marginal distributions with respect to Q can be obtained analytically in closed form. EP fixes the parameters of Q to approximately minimize the Kullback-Leibler divergence between the exact posterior and Q. As a side effect, EP also provides an estimate of the model evidence, which can be useful to perform model selection.

In many probabilistic models that assume i.i.d. observations the joint probability can be expressed as a product of several factors. In the specific case of a linear regression model, the joint probability of $\mathbf{w}$, $\mathbf{z}$ and $\mathbf{y}$ conditioned to $\mathbf{X}$ can be written as the product of 3 factors:

$$\mathcal{P}(\mathbf{w},\mathbf{z},\mathbf{y}|\mathbf{X}) = \mathcal{P}(\mathbf{y}|\mathbf{w},\mathbf{X})\mathcal{P}(\mathbf{w}|\mathbf{z})\mathcal{P}(\mathbf{z}) = \prod_{i=1}^{3} f_i(\mathbf{w},\mathbf{z})\,, \tag{8}$$

where the first factor corresponds to the likelihood, the second factor corresponds to the prior for $\mathbf{w}$, and the final factor corresponds to the prior for $\mathbf{z}$. Namely,

$$f_1(\mathbf{w},\mathbf{z}) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w},\sigma_0^2\mathbf{I})\,, \tag{9}$$

$$f_2(\mathbf{w},\mathbf{z}) = \prod_{j=1}^{d} z_{g(j)}\mathcal{N}(w_j|0,\sigma_{g(j)}^2) + (1-z_{g(j)})\delta(w_j)\,,$$

$$f_3(\mathbf{w},\mathbf{z}) = \prod_{g=1}^{G} p_{0,g}^{z_g}(1-p_{0,g})^{1-z_g}\,. \tag{10}$$

EP approximates each exact factor f_i by a simpler factor $\tilde{f}_i$ such that

$$\prod_{i=1}^{3} f_i(\mathbf{w},\mathbf{z}) \approx \prod_{i=1}^{3} \tilde{f}_i(\mathbf{w},\mathbf{z})\,.$$

All approximate factors $\tilde{f}_i$ are constrained to belong to the same family of exponential distributions, but they do not have to integrate to one. Once normalized with respect to $\mathbf{w}$, and $\mathbf{z}$, (8) becomes the

exact posterior distribution (5). Similarly, the normalized product of the $\tilde{f}_i$ becomes an approximation to the posterior:

$$Q(\mathbf{w},\mathbf{z}) = \frac{1}{Z}\prod_i \tilde{f}_i(\mathbf{w},\mathbf{z}), \tag{11}$$

where Z is the normalization constant which approximates $\mathcal{P}(\mathbf{y}|\mathbf{X})$. The exponential family of distributions is closed under the product operation. Therefore, Q has the same simple exponential form as the $\tilde{f}_i$ and Z can be readily computed. In practice, the form of Q is selected first, and the $\tilde{f}_i$ are then constrained to have the same form as Q. For each approximate factor $\tilde{f}_i$, one considers $Q^{\setminus i} \propto Q/\tilde{f}_i$. EP iteratively updates each $\tilde{f}_i$ one by one while minimizing the Kullback-Leibler (KL) divergence between $f_i Q^{\setminus i}$ and $\tilde{f}_i Q^{\setminus i}$. The KL-divergence minimized by EP includes a correction term so that it can be applied to un-normalized distributions (Zhu and Rohwer, 1995). More precisely, each EP update step minimizes

$$\mathrm{KL}\left(f_i Q^{\setminus i} || \tilde{f}_i Q^{\setminus i}\right) = \sum_{\mathbf{z}} \int \left[f_i Q^{\setminus i} \log \frac{f_i Q^{\setminus i}}{\tilde{f}_i Q^{\setminus i}} + \tilde{f}_i Q^{\setminus i} - f_i Q^{\setminus i} \right] d\mathbf{w}. \tag{12}$$

with respect to $\tilde{f}_i$. Note that we have omitted in (12) the dependencies of $\tilde{f}_i$, f_i and $Q^{\setminus i}$ on the parameters $\mathbf{w}$ and $\mathbf{z}$ to improve the readability. Specifically, EP involves the following steps:

1. Initialize all $\tilde{f}_i$ and Q to be uniform (non-informative).
2. Repeat until Q converges:
   (a) Select an $\tilde{f}_i$ to refine and compute $Q^{\setminus i} \propto Q/\tilde{f}_i$.
   (b) Update $\tilde{f}_i$ to minimize $\mathrm{KL}\left(f_i Q^{\setminus i} || \tilde{f}_i Q^{\setminus i}\right)$.
   (c) Update the approximation $Q^{\mathrm{new}} \propto \tilde{f}_i Q^{\setminus i}$.
3. Evaluate $Z \approx \mathcal{P}(\mathbf{y}|\mathbf{X})$ as the integral of the product of all the approximate factors.

The optimization problem in step 2-(b) is convex with a single global optimum. The solution to this problem is found by matching the expected values of the sufficient statistics under $f_i Q^{\setminus i}$ and $\tilde{f}_i Q^{\setminus i}$ (Bishop, 2006). EP is not guaranteed to converge globally but extensive empirical evidence shows that most of the times it converges to a fixed point (Minka, 2001). Non-convergence can be prevented by *damping* the EP updates (Minka and Lafferty, 2002). Damping is a standard procedure and consists in setting

$$\tilde{f}_i = [\tilde{f}_i^{\mathrm{new}}]^{\xi} [\tilde{f}_i^{\mathrm{old}}]^{1-\xi} \tag{13}$$

in step 2-(b), where $\tilde{f}_i^{\mathrm{new}}$ is the updated factor and $\tilde{f}_i^{\mathrm{old}}$ is the factor before the update. $\xi \in [0,1]$ is a parameter which controls the amount of damping. When $\xi = 1$, the standard EP update operation is recovered. When $\xi = 0$, no update of the approximate factors occurs. In our experiments we set $\xi = 0.9$ and progressively decay its value at each iteration of EP by 1%. Such a strategy offers good practical results and EP appears to always converge to a stationary solution. Finally, when compared to other approximate inference methods, such as Monte Carlo sampling or variational inference, EP has shown good overall performances (Minka, 2001). EP is also the preferred method for approximate inference in linear models with the standard spike-and-slab prior (Hernández-Lobato, 2010).

3.1 The Posterior Approximation

We approximate the posterior (5) using a parametric distribution that belongs to the exponential family

$$Q(\mathbf{w},\mathbf{z}) = \mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{V}) \prod_{g=1}^{G} \text{Bern}(z_g|\sigma(p_g)) \,, \tag{14}$$

where $\mathcal{N}(\cdot|\mathbf{m},\mathbf{V})$ denotes the probability density of a multivariate Gaussian with mean vector $\mathbf{m}$ and covariance matrix $\mathbf{V}$, and $\text{Bern}(\cdot|\sigma(p_g))$ denotes the probability mass function of a Bernoulli distribution with success probability $\sigma(p_g)$ and $\sigma(\cdot)$ is the sigmoid function:

$$\sigma(x) = \frac{1}{1+\exp\{-x\}} \,.$$

In (14), $\mathbf{p} = (p_1,\ldots,p_G)^{\text{T}}$, $\mathbf{m}$ and $\mathbf{V}$ are free parameters that have to be estimated by EP. The particular choice of (14) makes all the required computations tractable and offers good experimental results detailed in Section 7. The logistic function eliminates numerical under-flow or over-flow errors and simplifies the EP update operations, in a similar way as when EP is applied to a linear regression model with the standard spike-and-slab prior (Hernández-Lobato, 2010).

The approximate factors $\tilde{f}_i$ must have the same functional form as (14) but need not be normalized. Furthermore, the exact factor f_1 corresponding to the likelihood (2), only depends on $\mathbf{w}$. The Bernoulli part of $\tilde{f}_1$ can thus be removed:

$$\tilde{f}_1(\mathbf{w}) = \tilde{s}_1 \exp\left\{ -\frac{1}{2} (\mathbf{w}-\tilde{\mathbf{m}}_1)^{\text{T}} \tilde{\mathbf{V}}_1^{-1} (\mathbf{w}-\tilde{\mathbf{m}}_1)^{\text{T}} \right\} , \tag{15}$$

where $\tilde{\mathbf{m}}_1$, $\tilde{\mathbf{V}}_1$ and $\tilde{s}_1$ are free parameters to be estimated by EP. The second and third approximate factor $\tilde{f}_2$ and $\tilde{f}_3$ also have a special form because the corresponding exact factors f_2 and f_3 factorize with respect to each component of $\mathbf{w}$ and $\mathbf{z}$, respectively. Furthermore, f_3 is independent of $\mathbf{w}$ and its corresponding Gaussian part can be ignored. These approximate factors are hence defined as:

$$\tilde{f}_2(\mathbf{w},\mathbf{z}) = \tilde{s}_2 \left[\prod_{j=1}^{d} \exp\left\{ -\frac{1}{2\tilde{v}_{2,j}} (w_j - \tilde{m}_{2,j})^2 \right\} \text{Bern}(z_{g(j)}|\sigma(\tilde{p}_{2,j})) \right] , \tag{16}$$

$$\tilde{f}_3(\mathbf{z}) = \tilde{s}_3 \left[\prod_{g=1}^{G} \text{Bern}(z_g|\sigma(\tilde{p}_{3,g})) \right] , \tag{17}$$

where $\tilde{s}_2$, $\tilde{s}_3$, $\tilde{\mathbf{m}}_2 = (\tilde{m}_{2,1},\ldots,\tilde{m}_{2,d})^{\text{T}}$, $\tilde{\mathbf{v}}_2 = (\tilde{v}_{2,1},\ldots,\tilde{v}_{2,d})^{\text{T}}$, $\tilde{\mathbf{p}}_2 = (\tilde{p}_{2,1},\ldots,\tilde{p}_{2,d})^{\text{T}}$ and $\tilde{\mathbf{p}}_3 = (\tilde{p}_{3,1},\ldots,\tilde{p}_{3,G})^{\text{T}}$ are free parameters to be estimated by EP. The constants $\tilde{s}_1$, $\tilde{s}_2$ and $\tilde{s}_3$ are introduced to make sure that $\tilde{f}_i Q^{\setminus i}$ and the corresponding $f_i Q^{\setminus i}$ integrate up to the same value.

Once the different parameters of $\tilde{f}_1$, $\tilde{f}_2$ and $\tilde{f}_3$ have been estimated, the corresponding parameters $\mathbf{m}$, $\mathbf{V}$ and $\mathbf{p}$ of Q can be easily computed by using (11) and the closure property of the exponential family under the product operation. Namely, the product of two un-normalized Gaussian distributions is another un-normalized Gaussian distribution. Similarly, the product of two Bernoulli distributions is another Bernoulli distribution. The specific details of the product rules for Gaussian

and Bernoulli distributions are described in the Appendix of Hernández-Lobato (2009). The parameters of Q given each $\tilde{f}_i$ are obtained from applying those rules:

$$\mathbf{V} = \left(\tilde{\mathbf{V}}_1^{-1} + \mathbf{\Lambda}^{-1}\right)^{-1}, \tag{18}$$

$$\mathbf{m} = \mathbf{V}\left(\tilde{\mathbf{V}}_1^{-1}\tilde{\mathbf{m}}_1 + \mathbf{\Lambda}^{-1}\tilde{\mathbf{m}}_2\right), \tag{19}$$

$$p_g = \sum_{g(j)=g} \tilde{p}_{2,j} + \tilde{p}_{3,g}, \quad \text{for } g = 1,\ldots,G, \tag{20}$$

where $\mathbf{\Lambda} = \text{diag}(\tilde{v}_{2,1},\ldots,\tilde{v}_{2,d})$ and $\text{diag}(a_1,\ldots,a_d)$ denotes a diagonal matrix with elements $a_1,\ldots,a_d$ in the diagonal. At first glance, the computation of $\mathbf{V}$ according to (18) requires the inversion of a $d \times d$ matrix, with d equal to the dimensionality of the data. However, if $\tilde{\mathbf{V}}_1^{-1}$ is known and the regression problem satisfies $d \gg n$ (as for the regression problems we consider), the Woodbury formula provides a faster alternative with a computational cost of $O(n^2 d)$, where n is the number of observed samples (the specific details are given in the next section).

3.2 EP Update Operations

This section details how to update each approximate factor $\tilde{f}_1$, $\tilde{f}_2$ and $\tilde{f}_3$ according to the steps 2-(a), 2-(b) and 2-(c) of the EP algorithm. These operations involve minimizing the KL divergence between $f_i Q^{\setminus i}$ and $\tilde{f}_i Q^{\setminus i}$ with respect to $\tilde{f}_i$. This problem is convex with a single global minimum which is found by setting $\tilde{f}_i$ so that the expected values of the sufficient statistics under $\tilde{f}_i Q^{\setminus i}$ and $f_i Q^{\setminus i}$ match after normalization (Bishop, 2006). To simplify the notation, we only present here the operations for $\xi = 1$, that is, when there is no damping in the EP updates. Incorporating the effect of damping in these operations is straight-forward and is omitted. In particular, the natural parameters of each approximate factor become a convex combination of the natural parameters before and after each update operation, as derived from (13). However, as shown in this section, only the approximate factor $\tilde{f}_2$ needs to be updated. The optimal parameters of $\tilde{f}_1$ and $\tilde{f}_3$ can be computed exactly and these factors need not be updated by EP. Finally, we note that the parameters $\tilde{s}_1$, $\tilde{s}_2$ and $\tilde{s}_3$ of $\tilde{f}_1$, $\tilde{f}_2$ and $\tilde{f}_3$ are only needed to compute the approximation of the marginal likelihood in step 3 of the EP algorithm. Their computation can thus be delayed until EP converges as described below.

We now describe how to compute the parameters $\tilde{\mathbf{m}}_1$ and $\tilde{\mathbf{V}}_1$ of the first approximate factor $\tilde{f}_1$, which is fairly simple. In particular, we note that the corresponding exact factor f_1 has a Gaussian form with respect to $\mathbf{w}$, namely the likelihood of the observed data, as described in (9). Furthermore, the form of $\tilde{f}_1$ in (15) is also Gaussian. This means that the factor f_1 can be approximated exactly by EP, independently of the values of the other approximate factors $\tilde{f}_2$ and $\tilde{f}_3$. We only have to set $\tilde{f}_1 = f_1$, since both $\tilde{f}_1$ and f_1 have the same form. Consequently $\tilde{f}_1$ needs not be re-estimated by EP through the steps 2-(a) to 2-(c) but set equal to f_1 at the beginning of the EP algorithm and kept constant afterwards. The parameters of $\tilde{f}_1$ are:

$$\tilde{\mathbf{V}}_1^{-1} = \frac{1}{\sigma_0^2}\mathbf{X}^{\mathrm{T}}\mathbf{X}, \qquad \tilde{\mathbf{V}}_1^{-1}\tilde{\mathbf{m}}_1 = \frac{1}{\sigma_0^2}\mathbf{X}^{\mathrm{T}}\mathbf{y}. \tag{21}$$

Note that $\tilde{\mathbf{m}}_1$ is not uniquely defined in (21) whenever $d > n$. More precisely, if $d > n$ then $\tilde{\mathbf{V}}_1^{-1}$ is not full rank, $\tilde{\mathbf{V}}_1$ does not exist and the likelihood (2) is not strictly concave. Therefore, when $d > n$, there are infinitely many vectors $\tilde{\mathbf{m}}_1$ that can be used as potential solutions and must satisfy (21). For this reason, it is better in practice to define $\tilde{f}_1$ in terms of its natural parameters $\tilde{\mathbf{V}}_1^{-1}$ and

$\tilde{\mathbf{V}}_1^{-1}\tilde{\mathbf{m}}_1$. Namely,

$$\tilde{f}_1(\mathbf{w}) = \tilde{s}_1 \exp\left\{ -\frac{1}{2}\mathbf{w}^{\mathrm{T}}\tilde{\mathbf{V}}_1^{-1}\mathbf{w} + \mathbf{w}^{\mathrm{T}}\tilde{\mathbf{V}}_1^{-1}\tilde{\mathbf{m}}_1 \right\},$$

where the constant terms can be included in $\tilde{s}_1$.

The optimal parameters $\tilde{\mathbf{p}}_3$ of the third approximate factor $\tilde{f}_3$ can be found following the same reasoning. Specifically, the exact factor f_3 has the same form as the approximate factor $\tilde{f}_3$, that is, a product of G Bernoulli distributions, one for each component of $\mathbf{z}$. See (10) and (17). Thus $\tilde{f}_3$ is set equal to f_3 and not iteratively re-estimated through the EP algorithm. The corresponding parameters of $\tilde{f}_3$ are:

$$\tilde{p}_{3,g} = \sigma^{-1}(p_{0,g}), \quad \text{for } g = 1,\ldots,G, \tag{22}$$

where $\sigma^{-1}(\cdot)$ is the logit function, that is, the inverse of the sigmoid function. Namely,

$$\sigma^{-1}(x) = \log\frac{x}{1-x}.$$

In (22) $p_{0,g}$ is the prior probability of using the g-th group of features for prediction.

Updating the approximate factor $\tilde{f}_2$ is somewhat more complex and requires implementing all steps 2-(a) to 2-(c) of the EP algorithm. To simplify the computation, each of the d components of $\tilde{f}_2$ that appear in (16) is updated in parallel, as suggested by Gerven et al. (2009). For this, we factorize $\tilde{f}_2$ as follows:

$$\tilde{f}_2(\mathbf{w},\mathbf{z}) = \prod_{j=1}^{d} \tilde{s}_{2,j}\tilde{f}_{2,j}(w_j, z_{g(j)}), \tag{23}$$

where

$$\tilde{s}_2 = \prod_{j=1}^{d}\tilde{s}_{2,j}, \qquad \tilde{f}_{2,j}(w_j, z_{g(j)}) = \exp\left\{ -\frac{1}{2\tilde{v}_{2,j}}(w_j - \tilde{m}_{2,j})^2 \right\} \mathrm{Bern}(z_{g(j)}|\sigma(\tilde{p}_{2,j})).$$

Note that in (23) each component $\tilde{f}_{2,j}$ is the product of a univariate Gaussian distribution and a Bernoulli distribution. Thus, we only need the marginal distributions of Q for each component of $\mathbf{w}$ and $\mathbf{z}$ to obtain the corresponding update operations. The marginal distribution of Q for each component of $\mathbf{z}$, z_g, with $g = 1,\ldots,G$, is a Bernoulli distribution with probability parameter p_g, where p_g is defined in (20). Finding the means $\mathbf{m}$ and the variances $\mathrm{diag}(\mathbf{V})$ of each marginal distribution of Q for each component of $\mathbf{w}$ is more difficult. In principle, we could use (18) and (19) for computing $\mathrm{diag}(\mathbf{V})$ and $\mathbf{m}$, respectively. However, such a computation would require inverting a $d \times d$ matrix, where d is the number of dimensions. The Woodbury formula offers an efficient alternative when $d \gg n$, with a computational cost in $O(n^2 d)$:

$$\mathbf{V} = \mathbf{\Lambda} - \mathbf{\Lambda}\mathbf{X}^{\mathrm{T}}\left[\mathbf{I}\sigma_0^2 + \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{\mathrm{T}}\right]^{-1}\mathbf{X}\mathbf{\Lambda}, \tag{24}$$

where $\mathbf{\Lambda}$ is the diagonal matrix defined in (18) and $\mathbf{I}\sigma_0^2 + \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{\mathrm{T}}$ is a $n \times n$ matrix. Given this representation for $\mathbf{V}$, the value of the mean parameter $\mathbf{m}$ of Q can be computed using (19) and (21) in $O(n^2 d)$ steps. Namely,

$$\mathbf{m} = \mathbf{\Lambda}\boldsymbol{\eta} - \mathbf{\Lambda}\mathbf{X}^{\mathrm{T}}\left[\mathbf{I}\sigma_0^2 + \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{\mathrm{T}}\right]^{-1}\mathbf{X}\mathbf{\Lambda}\boldsymbol{\eta}, \tag{25}$$

where $\boldsymbol{\eta} = \mathbf{X}^T\mathbf{y}/\sigma_0^2 + \boldsymbol{\Lambda}^{-1}\tilde{\mathbf{m}}_2$ is a d-dimensional column vector. The $d \times d$ components of $\mathbf{V}$ need not be evaluated but only the diagonal of this matrix for the variances of the marginals. Consequently the cost of computing $\text{diag}(\mathbf{V}) = (V_{11}, \ldots, V_{dd})^{\text{T}}$ and $\mathbf{m}$ is in $O(n^2 d)$, which is linear in the dimensionality d of the regression problem. The use of the Woodbury formula may lead to numerical instabilities, which however have never been observed for the regression problems considered in Section 7.

Given $\text{diag}(\mathbf{V})$ and $\mathbf{m}$, we compute for each $\tilde{f}_{2,j}$ the corresponding parameters of the marginal distributions of w_j and $z_{g(j)}$ under $Q^{\setminus 2,j} \propto Q/\tilde{f}_{2,j}$. These parameters are obtained using the rules for the quotient between Gaussian distributions and the quotient between Bernoulli distributions. These rules are described in the Appendix of Hernández-Lobato (2009). Consider $m_j^{\setminus 2,j}$ and $v_j^{\setminus 2,j}$ to be respectively the mean and the variance of the marginal distribution of w_j under $Q^{\setminus 2,j}$. Similarly, let $\sigma(p_g^{\setminus 2,j})$, with $g = g(j)$, be the probability of using the g-th group for prediction under $Q^{\setminus 2,j}$. These parameters are obtained from $\text{diag}(\mathbf{V})$, $\mathbf{m}$ and the parameters of $\tilde{f}_{2,j}$ as follows:

$$
\begin{aligned}
v_j^{\setminus 2,j} &= \left(V_{jj}^{-1} - \tilde{v}_{2,j}^{-1}\right)^{-1}, \\
m_j^{\setminus 2,j} &= v_j^{\setminus 2,j}\left(V_{jj}^{-1} m_j - \tilde{v}_{2,j}^{-1}\tilde{m}_{2,j}\right), \\
p_g^{\setminus 2,j} &= p_g - \tilde{p}_{2,j},
\end{aligned}
\tag{26}
$$

where V_{jj} and m_j are respectively the variance and the mean of the marginal distribution of w_j under Q.

Once $Q^{\setminus 2,j}$ has been computed, we find the corresponding approximate factor $\tilde{f}_{2,j}$ which minimizes the KL divergence between $f_{2,j}Q^{\setminus 2,j}$ and $\tilde{f}_{2,j}Q^{\setminus 2,j}$, where $f_{2,j}$ is obtained from a factorization of f_2 equivalent to the one described for $\tilde{f}_2$ in (23). Specifically, $f_{2,j}(w_j, z_{g(j)}) = z_{g(j)}\mathcal{N}(w_j|0,\sigma_{g(j)}^2) + (1 - z_{g(j)})\delta(w_j)$. Given $Q^{\setminus 2,j}$ the optimal parameters of $\tilde{f}_{2,j}$ are:

$$
\begin{aligned}
\tilde{v}_{2,j}^{\text{new}} &= \frac{1}{a_j^2 - b_j} - v_j^{\setminus 2,j}, \\
\tilde{m}_{2,j}^{\text{new}} &= m_j^{\setminus 2,j} + \frac{a_j}{a_j^2 - b_j}, \\
\tilde{p}_{2,j}^{\text{new}} &= \log \mathcal{N}(0|m_j^{\setminus 2,j}, v_j^{\setminus 2,j} + v_0) - \log \mathcal{N}(0|m_j^{\setminus 2,j}, v_j^{\setminus 2,j}),
\end{aligned}
$$

where v_0 is the marginal variance of the slab and a_j and b_j are constants defined as:

$$
\begin{aligned}
a_j &= \sigma\left(\tilde{p}_{2,j}^{\text{new}} + p_g^{\setminus 2,j}\right)\frac{m_j^{\setminus 2,j}}{v_j^{\setminus 2,j} + v_0} + \sigma\left(-\tilde{p}_{2,j}^{\text{new}} - p_g^{\setminus 2,j}\right)\frac{m_j^{\setminus 2,j}}{v_j^{\setminus 2,j}}, \\
b_j &= \sigma\left(\tilde{p}_{2,j}^{\text{new}} + p_g^{\setminus 2,j}\right)\frac{(m_j^{\setminus 2,j})^2 - v_j^{\setminus 2,j} - v_0}{(v_j^{\setminus 2,j} + v_0)^2} + \sigma\left(-\tilde{p}_{2,j}^{\text{new}} - p_g^{\setminus 2,j}\right)\frac{(m_j^{\setminus 2,j})^2 - v_j^{\setminus 2,j}}{(v_j^{\setminus 2,j})^2}.
\end{aligned}
$$

These update operations require the value of $v_j^{\setminus 2,j}$ to be positive. In some rare situations a negative value is found for $v_j^{\setminus 2,j}$ after removing $\tilde{f}_{2,j}$ from the posterior approximation Q to compute $Q^{\setminus 2,j}$. In such a rare occurrence, the corresponding update of $\tilde{f}_{2,j}$ is not performed. Furthermore, a negative value for $\tilde{v}_{2,j}^{\text{new}}$ may be observed when computing the optimal parameters for $\tilde{f}_{2,j}$. Negative

variances are common in many EP implementations (Minka, 2001; Minka and Lafferty, 2002). In this case, the approximate factors are not un-normalized density functions, but correction factors that compensate for errors in the first approximate factor $\tilde{f}_1$. It has been observed in the literature that negative variances can lead to erratic behavior in EP and to longer convergence times (Seeger, 2008; Hernández-Lobato, 2010). To avoid these problems, we minimize the KL between $f_{2,j}Q^{\backslash 2,j}$ and $\tilde{f}_{2,j}Q^{\backslash 2,j}$ with the constraint of $\tilde{v}_{2,j}^{\text{new}}$ being positive. In this case, whenever the optimal $\tilde{v}_{2,j}^{\text{new}}$ is negative, we simply set $\tilde{v}_{2,j}^{\text{new}} = v_\infty$, where v_∞ is a large positive constant. The update of the other parameters of $\tilde{f}_{2,j}$ ($\tilde{m}_{2,j}$ and $\tilde{p}_{2,j}$) is kept unchanged. This approach, already used in the linear regression model with standard spike-and-slab priors described by Hernández-Lobato (2010), offered improved convergence results.

Once all the approximate factors $\tilde{f}_{2,j}$, with $j = 1,\ldots,d$, have been updated in parallel, Q needs to be recomputed as the product of all the approximate factors. This corresponds to step 2-(c) of the EP algorithm. For this, we can use (18), (19) and (20). However, we have already described how to obtain the means $\mathbf{m}$ and the variances $\text{diag}(\mathbf{V})$ of the marginals of Q for $\mathbf{w}$ to update each approximate factor $\tilde{f}_{2,j}$. Thus, in practice, one only has to use (20) to recompute the parameter $\mathbf{p}$ of Q, which is needed in (26). There is no need to recompute the complete covariance matrix $\mathbf{V}$ of the Gaussian part of Q since only the diagonal of this matrix is strictly needed for the EP updates. In summary, the total cost of the EP algorithm under the assumption of a constant number of iterations until convergence is in $O(n^2 d)$, where n is the number of training instances and d is the number of features.

3.3 Approximation of the Model Evidence

The Bayesian approach to model selection specifies that the model with the largest evidence should be preferred (Bishop, 2006; MacKay, 2003). The model evidence is defined in (5) as $\mathcal{P}(\mathbf{y}|\mathbf{X})$, that is, the normalization constant used to compute the posterior distribution from the joint distribution of the model parameters and the data. It can also be described as the probability that the targets $\mathbf{y}$ are generated from the design matrix $\mathbf{X}$ using a linear model whose coefficient vector $\mathbf{w}$ is randomly sampled from the assumed prior distribution. The model evidence naturally achieves a balance between penalizing model complexity and rewarding models that provide a good fit to the training data (Bishop, 2006). However, a practical difficulty in computing $\mathcal{P}(\mathbf{y}|\mathbf{X})$ is its expensive computational cost. Specifically, the exact evaluation of $\mathcal{P}(\mathbf{y}|\mathbf{X})$ is infeasible for large G since it involves a sum of 2^G terms. These are the 2^G different configurations for the vector of latent variables $\mathbf{z}$. Nevertheless, if needed, EP can be used to efficiently compute an approximation once it has converged, as described in step 3 of the algorithm:

$$\mathcal{P}(\mathbf{y}|\mathbf{X}) \approx \int \sum_{\mathbf{z}} \tilde{f}_1(\mathbf{w},\mathbf{z})\tilde{f}_2(\mathbf{w},\mathbf{z})\tilde{f}_3(\mathbf{w},\mathbf{z})d\mathbf{w}\,. \tag{27}$$

Since all the approximate factors $\tilde{f}_1$, $\tilde{f}_2$ and $\tilde{f}_3$ have simple exponential forms, (27) can be readily evaluated. We only have to use the formulas for the product of Gaussian and Bernoulli distributions. These formulas are described in the Appendix of Hernández-Lobato (2009). The evaluation of (27) also requires the computation of the parameters $\tilde{s}_1$, $\tilde{s}_2$ and $\tilde{s}_3$ of each approximate factor. These parameters are estimated once EP has converged and their specific values are fixed to guarantee that

$\tilde{f}_i Q^{\setminus i}$ and the corresponding $f_i Q^{\setminus i}$ integrate up to the same value:

$$\begin{aligned}
\log(\tilde{s}_1) &= -\frac{n}{2}\log(2\pi\sigma_0^2) - \frac{1}{2\sigma_0^2}\mathbf{y}^{\mathrm{T}}\mathbf{y}\,, \\
\log(\tilde{s}_{2,j}) &= \log(\kappa_j) - \log\left(\sigma(\tilde{p}_{2,j})\sigma(p_{g(j)}^{\setminus 2,j}) + \sigma(-\tilde{p}_{2,j})\sigma(-p_{g(j)}^{\setminus 2,j})\right) - \frac{1}{2}\log\left(2\pi\tilde{v}_{2,j}\right) \\
&\quad - \log\left(\mathcal{N}(0|m_j^{\setminus 2,j} - \tilde{m}_{2,j}, \tilde{v}_{2,j} + V_{jj}^{\setminus 2,j})\right)\,, \quad \text{for } j = 1,\ldots,d \\
\log(\tilde{s}_3) &= 0\,,
\end{aligned}$$

were $\kappa_j = \sigma(p_{g(j)}^{\setminus 2,j})\mathcal{N}(0|m_j^{\setminus 2,j}, v_0 + V_{jj}^{\setminus 2,j}) + \sigma(-p_{g(j)}^{\setminus 2,j})\mathcal{N}(0|m_j^{\setminus 2,j}, V_{jj}^{\setminus 2,j})$ and $\tilde{s}_2 = \prod_{j=1}^d \tilde{s}_{2,j}$ since we have further factorized $\tilde{f}_2$ as the product of d factors $\tilde{f}_{2,j}$ with $j = 1,\ldots,d$. See (23) for further details.

Given these values, we can approximate the logarithm of the model evidence by

$$\begin{aligned}
\log \mathcal{P}(\mathbf{y}|\mathbf{X}) \approx \sum_{i=1}^{3}\log\tilde{s}_i + \sum_{g=1}^{G}\log\left(\sigma(\tilde{p}_{3,g})\prod_{g(j)=g}\sigma(\tilde{p}_{2,j}) + \sigma(-\tilde{p}_{3,g})\prod_{g(j)=g}\sigma(-\tilde{p}_{2,j})\right) \\
- \frac{1}{2}\tilde{\mathbf{m}}_2^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\tilde{\mathbf{m}}_2 + \frac{d}{2}\log(2\pi) + \frac{1}{2}\log(|\mathbf{V}|) + \frac{1}{2}\boldsymbol{\upsilon}^{\mathrm{T}}\mathbf{V}\boldsymbol{\upsilon}\,, \qquad (28)
\end{aligned}$$

where $\mathbf{V}$ is the covariance matrix of the Gaussian part of Q and $\boldsymbol{\upsilon}$ is a d-dimensional vector defined as $\boldsymbol{\upsilon} = 1/\sigma_0^2\mathbf{X}^{\mathrm{T}}\mathbf{y} + \boldsymbol{\Lambda}^{-1}\tilde{\mathbf{m}}_2$. In practice it is better to work with the logarithm of the approximation to $\mathcal{P}(\mathbf{y}|\mathbf{X})$ to avoid numerical over-flow and under-flow errors. Furthermore, the computation of $|\mathbf{V}|$ can be efficiently implemented when $d \gg n$ using the Sylvester's determinant formula and $\boldsymbol{\upsilon}^{\mathrm{T}}\mathbf{V}\boldsymbol{\upsilon}$ can be evaluated in $O(n^2 d)$ steps using the representation for $\mathbf{V}$ given in (24).

We note that one should use the model evidence with care to perform model selection. In particular, if the assumptions made about the form of the model are not accurate enough, the results obtained by Bayesian model comparison can be misleading, as indicated by Bishop (2006). This is precisely the case of the experiments reported in Section 7.2, where we have observed that the approximation to the model evidence (28) provides sub-optimal decisions to choose the model hyper-parameters p_0 and v_0. Thus, in a practical application it is wise to keep aside an independent validation set to evaluate the overall performance of the final system.

3.4 Prediction and Identification of Relevant Groups

Once EP has converged and Q has been estimated, we can use this approximation for making predictions. In particular, we only have to substitute the posterior approximation in (6) to obtain an approximate predictive distribution for the target y_{new} associated to a new instance $\mathbf{x}_{\text{new}}$:

$$\begin{aligned}
\mathcal{P}(y_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{y}, \mathbf{X}) &\approx \sum_{\mathbf{z}}\int \mathcal{P}(y_{\text{new}}|\mathbf{w}, \mathbf{x}_{\text{new}})\,Q(\mathbf{w}, \mathbf{z})\,d\mathbf{w} \\
&= \mathcal{N}(y_{\text{new}}|\mathbf{m}^{\mathrm{T}}\mathbf{x}_{\text{new}}, \mathbf{x}_{\text{new}}^{\mathrm{T}}\mathbf{V}\mathbf{x}_{\text{new}} + \sigma_0^2)\,, \qquad (29)
\end{aligned}$$

where $\mathbf{m}$ and $\mathbf{V}$ are the mean vector and the covariance matrix of the posterior approximation Q for $\mathbf{w}$. Both $\mathbf{x}_{\text{new}}^{\mathrm{T}}\mathbf{V}\mathbf{x}_{\text{new}}$ and $\mathbf{m}^{\mathrm{T}}\mathbf{x}_{\text{new}}$ can be efficiently computed in $O(n^2 d)$ steps using the representations given in (24) and (25), respectively. Finally, if one is only interested in the expected value of y_{new} and not in the uncertainty of the prediction, the computation of $\mathbf{x}_{\text{new}}^{\mathrm{T}}\mathbf{V}\mathbf{x}_{\text{new}}$ can be omitted.

The posterior approximation Q is also very useful to identify the groups of features that are more relevant for prediction after substituting the exact posterior by Q in (7). Namely,

$$\mathcal{P}(\mathbf{z}|\mathbf{y},\mathbf{X}) \approx \int Q(\mathbf{w},\mathbf{z})\, d\mathbf{w} = \prod_{g=1}^{G} \text{Bern}(z_g|\sigma(p_g))\,,$$

where $\sigma(p_g)$ approximates the posterior probability of using the g-th group of features for prediction. Thus, we can use the parameters $p_1,\ldots,p_G$ of Q to identify the groups which are more likely used for prediction. More precisely, we should observe a bi-separation of the different groups in two sets according to these parameters. The first set would contain features which are unlikely to be used for prediction. By contrast, the second set would contain features which are used for prediction with high posterior probability.

4. Sequential Experimental Design

Sequential experimental design (also known in the literature as optimal design or active learning) deals with the problem of saving on expensive experiments to obtain the highest level of information about the different latent variables or parameters of the assumed model (Seeger, 2008; Chaloner and Verdinelli, 1995; Fedorov, 1972). In our particular scenario, sequential experimental design tries to answer the following problem. Consider several candidate points $\mathbf{x}_{\text{new}}$ that are available for inclusion into the training set of the model. At which of these points should the corresponding target value y_{new} be sampled to obtain as much new information as possible about the unknown $\mathbf{w}$? Assuming that $\mathbf{x}_{\text{new}}$ and y_{new} are both known, Seeger (2008) and MacKay (1991) describe some natural scores that can be used to answer this question: for example, the decrease in the posterior uncertainty or the gain in information from the current posterior distribution $\mathcal{P}(\mathbf{w}|\mathbf{y},\mathbf{X})$ to the updated posterior distribution $\mathcal{P}'(\mathbf{w}|\mathbf{y},\mathbf{X})$ that is obtained once the new data instance is introduced in the training set. Information gain can be measured in terms of the KL divergence between these two distributions. Namely, we could aim to maximize $\text{KL}(\mathcal{P}'||\mathcal{P})$. The target value y_{new} is however often not available for the candidate points $\mathbf{x}_{\text{new}}$. A natural alternative is obtained by marginalizing this score over the expected distribution of y_{new} given the assumed model. Such a score is defined as $\mathbb{E}\{\text{KL}(\mathcal{P}'||\mathcal{P})\}$, where the expectation goes over $\mathcal{P}(y_{\text{new}}|\mathbf{x}_{\text{new}},\mathbf{y},\mathbf{X})$, that is, the predictive distribution for y_{new} given the current model. Another potential score to be used for this purpose is the expected decrease in the entropy of the posterior distribution once the new instance has been included in the training set. That is, $\mathbb{E}\{H[\mathcal{P}] - H[\mathcal{P}']\}$, where $H[\mathcal{P}]$ and $H[\mathcal{P}']$ respectively denote the entropy of the posterior distribution before and after the inclusion of the candidate point $\mathbf{x}_{\text{new}}$ into the training set. MacKay (1991) actually shows that both scores are equivalent and lead to the selection of the same instance $\mathbf{x}_{\text{new}}$. Consequently we focus on this last score in the rest of this section. Such score has also been used by Ji and Carin (2007) and by Seeger (2008) to perform sequential experiment design using a sparse linear model and by Lawrence et al. (2003) in the context of sparse Gaussian processes.

As discussed in Section 2.1, the computation of the exact posterior distribution $\mathcal{P}(\mathbf{w}|\mathbf{y},\mathbf{X})$ is intractable in practice. Thus, we have to resort to the EP posterior approximation Q for the estimation of the entropy of the posterior distribution. More precisely, we replace the entropy of the exact posterior $H[\mathcal{P}]$ by the entropy of the EP approximation $H[Q]$. The score to maximize is now defined as $\mathbb{E}\{H[Q] - H[Q']\}$, where $H[Q]$ and $H[Q']$ respectively denote the entropy of Q before and after the inclusion of the candidate point $\mathbf{x}_{\text{new}}$ into the training set.

Assume we would like to score a new instance $\mathbf{x}_{\text{new}}$. Further consider that Q has been marginalized over the latent variables $\mathbf{z}$, which means that only the Gaussian part of Q remains. Under these assumptions, the logarithm of the entropy of Q is:

$$\log H[Q] = -\frac{1}{2}\log\left(\left|\frac{1}{\sigma_0^2}\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{\Lambda}^{-1}\right|\right) + C \tag{30}$$

where C summarizes some constants that are independent of $\mathbf{x}_{\text{new}}$, and $\mathbf{\Lambda}$ is the diagonal matrix defined in (20). Once $\mathbf{x}_{\text{new}}$ has been included in the training set, the logarithm of the entropy of the updated posterior approximation Q' is:

$$\log H[Q'] = \log H[Q] - \frac{1}{2}\log\left(1 + \frac{1}{\sigma_0^2}\mathbf{x}_{\text{new}}^{\mathrm{T}}\mathbf{V}\mathbf{x}_{\text{new}}\right) + C\,. \tag{31}$$

where we have used the Sylvester's determinant theorem and $\mathbf{V}$ is the covariance matrix of the Gaussian part of Q, that is, the posterior approximation before the update. Furthermore, in (31) we have made the assumption that the parameters $\tilde{\mathbf{v}}_2$ of the approximate factor $\tilde{f}_2$ in Q (the diagonal entries of the matrix $\mathbf{\Lambda}$) are constant when the candidate point $\mathbf{x}_{\text{new}}$ is included in the training set. Of course, this assumption need not be satisfied in practice and EP has to be run to find the updated parameters $\tilde{\mathbf{v}}_2$ of $\tilde{f}_2$ in Q'. However, running the EP algorithm each time a candidate point has to be scored would be very expensive. A simple alternative, used in (31), keeps the approximate factor $\tilde{f}_2$ constant and lets $\tilde{f}_1$ vary in Q'. In other words, for the purpose of scoring new candidate points we treat the model as purely linear-Gaussian. Doing so lets us compute the score very efficiently and many candidate points can be evaluated. The same approximation has been used by Seeger (2008) in a linear regression model with Laplace priors. Note that (30) does not depend on the target y_{new} associated to $\mathbf{x}_{\text{new}}$. This means that $H[Q]$ is independent of y_{new} and can be taken out from the expectation with respect to $\mathcal{P}(y_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{y}, \mathbf{X})$. The same applies to $H[Q']$ under the assumptions described. Thus, for the purpose of scoring candidate points, we can ignore the expectation over $\mathcal{P}(y_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{y}, \mathbf{X})$.

Since $H[Q]$ is constant for all candidate points to be scored, expression (31) indicates that the candidate point $\mathbf{x}_{\text{new}}$ to be included in the model is the one that maximizes $\mathbf{x}_{\text{new}}^{\mathrm{T}}\mathbf{V}\mathbf{x}_{\text{new}}$. This is precisely the term which specifies the uncertainty in the prediction of the target value y_{new} associated to the instance $\mathbf{x}_{\text{new}}$ under the current model. See (29) for further details. In other words, those points for which the model is the most unsure about their target value are preferred. These points are expected to be the most informative. However, in practice, the candidate points can also be unknown. In such a situation, the optimal candidate point $\mathbf{x}_{\text{new}}$ to include in the model is a vector parallel to the eigenvector of the covariance matrix $\mathbf{V}$ with the largest associated eigenvalue. Such vector can be efficiently found, for example, using the power method. The total computational cost of the power method under the assumption of a constant number of iterations until convergence is $O(n^2 d)$ when $d \gg n$. For guaranteeing such computational cost, the algorithm must be efficiently implemented using the representation for $\mathbf{V}$ given in (24). Once $\mathbf{x}_{\text{new}}$ has been found, we can carry out the required experiments to measure the associated target value y_{new}. This is precisely the procedure which is followed in adaptive compressed sensing experiments, where a sparse signal (the model coefficients $\mathbf{w}$) is reconstructed from a small number of sequentially designed measurements (Seeger, 2008; Ji and Carin, 2007). The model described here also includes prior knowledge about groups of components of the sparse signal which are expected to be jointly equal to zero or jointly different from zero. Section 7 shows that the inclusion of this prior knowledge in the inductive procedure leads to improved reconstruction errors.

5. Related Methods for Group Feature Selection

In this section we review other methods that are available in the literature to perform feature selection at the group level. Some of these techniques involve a complete Bayesian approach, similar to the one described in this document, but they use Markov chain Monte Carlo sampling techniques instead of EP to approximate the posterior distribution of $\mathbf{w}$ given the observed data (Bishop, 2006; MacKay, 2003). Other techniques do not necessarily proceed this way and specify instead a particular objective function which is optimized. This function includes an error loss and a set of constraints to enforce sparsity at the group level. Finally, other techniques follow a type-II maximum likelihood approach (Bishop, 2006) in which the model evidence is optimized with respect to some hyper-parameters to enforce sparsity at the group level.

5.1 The Group LASSO

We start by reviewing the group LASSO (Yuan and Lin, 2006; Kim et al., 2006), which is probably the most popular method employed for group feature selection. This method is a natural extension of the LASSO (Tibshirani, 1996) and consists in estimating a linear predictor by minimizing a squared loss error function evaluated on the observed data, under a series of constraints which enforce sparsity at the group level. Unlike the approach described in this document, this method does not provide a posterior probability distribution for $\mathbf{w}$, but a point estimate. Specifically, the estimator for $\mathbf{w}$ in the group LASSO is:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \quad \mathcal{L}(\mathbf{w}) \qquad\qquad \text{s.t.} \quad \sum_{g=1}^{G} s(d_g)||\mathbf{w}_g|| \leq k\,, \tag{32}$$

where $\mathcal{L}$ is a convex loss function evaluated on the training data, for example, the squared loss defined as $\sum_{i=1}^{n}(y_i - \mathbf{w}^{\mathrm{T}}\mathbf{x}_i)^2$; $\mathbf{w}_g$ is a vector that contains the components of $\mathbf{w}$ within the g-th group; d_g is the dimension of the vector $\mathbf{w}_g$; $s(\cdot)$ is a scaling function used to account for groups of different sizes; $||\mathbf{w}_g||$ is a norm of the vector $\mathbf{w}_g$; and k is a positive regularization parameter. Besides the squared loss, other authors have also considered a logistic regression loss to address classification problems (Meier et al., 2008). The norm that penalizes each vector $\mathbf{w}_g$ is typically the ℓ_2-norm, although the ℓ_∞-norm has also been considered by Vogt and Roth (2010). The function $s(\cdot)$ is often set to be the square root (Meier et al., 2008). The group LASSO has been shown to be asymptotically consistent under certain conditions (Bach, 2008; Meier et al., 2008). However, when $d \gg n$, the minimizer of (32) may not be unique (Roth and Fischer, 2008; Vogt and Roth, 2010). The level of sparsity in the group LASSO is determined by the regularization parameter k. The smaller the value of k the sparser the solution at the group level and vice-versa. The optimal value of k is specific to the problem under consideration. Typically, it is fixed by minimizing an independent estimate of the generalization error obtained by cross-validation. An optimal solution to (32) can be obtained using the efficient algorithm described by Roth and Fischer (2008). Finally, there exists an equivalent formulation of the group LASSO where the optimization problem is un-constrained, but the loss function is penalized by the sum of the ℓ_2-norms of the group components of $\mathbf{w}$ (Yuan and Lin, 2006). Namely, (32) is equivalent to

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \quad \mathcal{L}(\mathbf{w}) + \gamma \sum_{g=1}^{G} s(d_g)||\mathbf{w}_g||\,. \tag{33}$$

for some $\gamma > 0$, which plays an opposite role to k in (32). In particular, the larger the value of γ, the sparser the solution is at the group level.

The group LASSO has proven to be useful in many domain applications (Kim et al., 2006; Meier et al., 2008; Roth and Fischer, 2008). Nevertheless, this method suffers from the problem of finding meaningful variance and covariance estimates for the regression coefficients $\mathbf{w}$, as described by Raman et al. (2009). These estimates could be easily obtained by using the formulation given in (33) to compute the Hessian at the optimal solution. This is, for example, the approach followed by the Laplace approximation to perform approximate Bayesian inference (MacKay, 2003). Unfortunately, the objective function in (33) is not differentiable at the optimum as a consequence of the regularization term which enforces sparsity at the group level. This means that the Hessian is un-defined at the optimum and the variance and covariance estimates of $\mathbf{w}$ cannot be computed in practice. In the case of the model described in this document these estimates can be very useful, as reported in Section 7, to perform sequential experimental design. Finally, another drawback of the group LASSO is that, as such, it does not allow to favor the selection of specific groups of features that are *a priori* believed to be more relevant. In the model described in this document, the inclusion of this type of prior knowledge is very easy by specifying different values for $p_{0,g}$, with $g = 1, \ldots, G$, in the prior distribution (4) for $\mathbf{z}$.

5.2 The Bayesian Group LASSO

The Bayesian group LASSO is proposed in Raman et al. (2009) as a full Bayesian treatment of the group LASSO to overcome the problem of covariance estimation just described. From a probabilistic perspective the group LASSO can be understood as a standard linear regression model with Gaussian noise and a product of multi-variate Laplacian priors over the regression coefficients. In particular, the target values $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ are assumed to be generated according to $y_i = \mathbf{w}^{\mathrm{T}}\mathbf{x}_i + \varepsilon_i$, with $\varepsilon_i \sim \mathcal{N}(0, \sigma_0^2)$, which implies a Gaussian likelihood equivalent to the one described in (2). Assume d_g is the dimension of the g-th group of features. Consider now for each group of coefficients $\mathbf{w}_g$, with $g = 1, \ldots, G$, a multivariate and spherical d_g-dimensional Multi-Laplace prior, which can be expressed as a hierarchical normal-gamma model. Namely,

$$\begin{aligned}\mathcal{P}(\mathbf{w}_g) &= \int \mathcal{N}(\mathbf{w}_g|\mathbf{0}, \lambda_g^2\mathbf{I})\mathrm{Gamma}\left(\lambda_g^2 \middle| \frac{d_g+1}{2}, \frac{2}{d_g\gamma^2}\right) d\lambda_g^2 \\ &\propto \left(d_g\gamma^2\right)^{\frac{d_g}{2}} \exp\left\{-\gamma\sqrt{d_g}||\mathbf{w}_g||_2\right\},\end{aligned}$$

where γ is a parameter which determines the degree of group sparsity, $||\cdot||_2$ represents the ℓ_2-norm, λ_g^2 can be seen as some latent parameter and $\mathrm{Gamma}(\cdot|a,b)$ denotes a gamma distribution with shape and scale parameters a and b, respectively. The complete prior for $\mathbf{w}$ is hence defined as $\mathcal{P}(\mathbf{w}) = \prod_{g=1}^{G} \mathcal{P}(\mathbf{w}_g)$. Consider now the posterior distribution of $\mathbf{w}$ under this likelihood and this prior distribution $\mathcal{P}(\mathbf{w}|\mathbf{y}) \propto \mathcal{P}(\mathbf{y}|\mathbf{X}, \mathbf{w}) \prod_{g=1}^{G} \mathcal{P}(\mathbf{w}_g)$. If we set $\sigma_0^2 = 1/2$, the group LASSO, as defined in (33), is obtained by maximizing the logarithm of $\mathcal{P}(\mathbf{w}|\mathbf{y})$ with respect to $\mathbf{w}$.

Instead of considering only a single point estimate of $\mathbf{w}$, the Bayesian group LASSO considers the complete posterior distribution for $\mathbf{w}$ given the observed data, under the model just described. Unfortunately, the exact computation of this distribution is intractable and closed form expressions to describe it cannot be obtained. This means that in practice one has to use approximate inference techniques. Raman et al. (2009) have proposed to use Markov chain Monte Carlo methods for this

purpose. In particular, a Gibbs sampling approach which iteratively generates samples from the conditional distributions of each coefficient w_j and each latent parameter λ_g^2 from the normal-gamma prior. In this document we consider a more efficient Gibbs sampling algorithm for the Bayesian group LASSO where the model coefficients are marginalized out and we directly sample from the unconditional distribution of the latent parameters λ_g^2, with $g = 1, \ldots, G$. Once these samples have been generated, we then sample from the conditional distribution of $\mathbf{w}$ given the latent parameters. Appendix A describes the details of this algorithm. Finally, even though it successfully provides covariance estimates, the Bayesian group LASSO does not consider favoring the selection of specific groups of features.

5.3 The Group Horseshoe

The group horseshoe introduced in this section is a natural extension of the robust horseshoe prior initially proposed to address sparse supervised learning problems (Carvalho et al., 2009). A model incorporating the horseshoe prior can be described by assuming Gaussian noise around the target values $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, as in the Bayesian model considered here and as in the Bayesian group LASSO. The likelihood for $\mathbf{y}$ given $\mathbf{w}$, $\mathcal{P}(\mathbf{y}|\mathbf{X}, \mathbf{w})$, is equivalent to (2). Under the horseshoe prior each component of $\mathbf{w}$, w_j, is assumed to be conditionally independent with a density which can be represented as a scale mixture of normals:

$$\mathcal{P}(w_j|\lambda_j, \tau) = \int \mathcal{N}(w_j|0, \lambda_j^2\tau^2)\mathcal{C}^+(\lambda_j|0,1)d\lambda_j\,, \tag{34}$$

where $\mathcal{C}^+(\cdot|0,1)$ is a half-Cauchy distribution with location and scale parameters equal to 0 and 1, respectively; λ_j^2 is a latent parameter; and τ is a shrinkage parameter which determines the level of sparsity: the smaller the value of τ, the sparser the prior. We note that (34) describes a hierarchical normal-half-Cauchy model. This prior has two interesting properties which make it useful for induction under the sparsity assumption for $\mathbf{w}$. First, Cauchy-like tails allow for large values of w_j. Second, it has an infinitely tall spike at the origin which favors values of w_j close to zero. A detailed analysis of this prior and several benchmark experiments which consider different regression problems illustrate its advantages with respect to other approaches for sparse learning (Carvalho et al., 2009).

The prior described in (34) can be easily generalized to address sparsity at the group level. For this, we only have to assume the same latent parameter λ_j for several components of $\mathbf{w}$. Specifically, we consider for each group of coefficients $\mathbf{w}_g$, with $g = 1, \ldots, G$, a multivariate and spherical d_g-dimensional prior, which can be expressed as a hierarchical normal-half-Cauchy model, as in (34):

$$\mathcal{P}(\mathbf{w}_g|\lambda_g, \tau) = \int \mathcal{N}(\mathbf{w}_g|\mathbf{0}, \lambda_g^2\tau^2\mathbf{I})\mathcal{C}^+(\lambda_g|0,1)d\lambda_g\,, \tag{35}$$

where λ_g is a latent parameter specific to each group and τ is a shrinkage parameter. The resulting prior has similar properties to the one-dimensional prior described in (34). That is, Cauchy-like tails to allow for large values of each component of $\mathbf{w}_g$ and an infinitely tall spike at the origin which favors values where all the components of $\mathbf{w}_g$ are close to zero. See Section 6 for further details on this prior.

The posterior distribution of $\mathbf{w}$ under the assumed Gaussian likelihood and the prior distribution introduced in (35) is given by $\mathcal{P}(\mathbf{w}|\mathbf{y}) \propto \mathcal{P}(\mathbf{y}|\mathbf{X}, \mathbf{w}) \prod_{g=1}^{G} \mathcal{P}(\mathbf{w}_g)$. As in the Bayesian group LASSO,

the exact computation of this distribution is intractable and, in practice, we have to resort to approximate techniques. Initially, one can think about using EP for this task. However, the specific application of the EP algorithm to this model is challenging. Specifically, the prior distribution suggested in (35) can not be evaluated exactly since it is not possible to evaluate the corresponding integral in closed form. Furthermore, this prior does not have a closed form convolution with the Gaussian distribution. This makes all the computations required by EP very difficult. Consequently we use a simpler alternative in this document. Namely, a Gibbs sampling technique similar to the one used for the Bayesian group LASSO. Appendix A explains the details of this algorithm.

5.4 Automatic Relevance Determination for Groups of Features

Another technique which can be used to perform feature selection at the group level is Automatic Relevance Determination (ARD) (Tipping, 2001; Ji et al., 2009). Like the other models described so far in this section, ARD also assumes a Gaussian likelihood for $\mathbf{y}$ given $\mathbf{w}$. Then, a zero mean factorizing Gaussian prior is fixed for $\mathbf{w}$. In the simplest formulation of ARD, this prior distribution has a different hyper-parameter α_j for each dimension of the problem (Li et al., 2002). Namely, $\mathcal{P}(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{A}^{-1})$, where $\mathbf{A}$ is a diagonal matrix with entries $A_{jj} = \alpha_j$, for $j = 1, \ldots, d$. Thus, each hyper-parameter α_j, is the inverse of the prior variance for the corresponding component w_j of $\mathbf{w}$. Since both the likelihood and the prior are Gaussian under this formulation, the model evidence can be evaluated exactly. This evidence can be maximized component-wise with respect to the hyper-parameters α_j, with $j = 1, \ldots, d$, using the fast algorithm suggested by Tipping and Faul (2003). Specifically, these authors provide a closed form solution for the optimal α_j while the other hyper-parameters are kept fixed. This means that one only has to iteratively optimize the model evidence with respect to each α_j until convergence. In such case, one typically finds that most of these hyper-parameters are driven to infinity during the optimization process. Consequently the posterior distribution of the coefficients of $\mathbf{w}$ corresponding to these hyper-parameters is set to a delta function centered at zero. Thus, this procedure can be used to induce $\mathbf{w}$ under the sparsity assumption.

Following the ARD principle sparsity at the group level can be easily obtained by considering a different hyper-parameter for each group of coefficients. Specifically, Ji et al. (2009) consider the following Gaussian prior distribution for $\mathbf{w}$:

$$\mathcal{P}(\mathbf{w}) = \prod_{g=1}^{G} \mathcal{P}(\mathbf{w}_g) = \prod_{g=1}^{G} \mathcal{N}(\mathbf{w}_g|\mathbf{0}, \alpha_g^{-1}\mathbf{I}),$$

where α_g is the inverse of the prior variances for each component of $\mathbf{w}_g$, that is, the vector of model coefficients within the g-th group. Given this prior distribution and a Gaussian likelihood such as the one described in (2), the corresponding model evidence is $\mathcal{P}(\mathbf{y}|\mathbf{X}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{C})$, where $\mathbf{C}$ is a $n \times n$ matrix defined as $\mathbf{C} = \sigma_0^2\mathbf{I} + \mathbf{X}\mathbf{A}^{-1}\mathbf{X}^{\mathrm{T}}$, and $\mathbf{A}$ is a diagonal matrix with components $A_{jj} = \alpha_g$, if the j-th feature belongs to the g-th group, for $j = 1, \ldots, d$. This value, $\mathcal{P}(\mathbf{y}|\mathbf{X})$, can be easily optimized iteratively with respect to each α_g using an algorithm similar to the one described by Tipping and Faul (2003). In this case one typically finds that most of the α_g's tend to infinity, enforcing the posterior for the corresponding $\mathbf{w}_g$ to be a spike at the origin. Unfortunately the optimization process is more difficult and there is no closed form solution for the optimal α_g while the other hyper-parameters are kept fixed. Ji et al. (2009) provide a closed form approximate solution which is shown to perform well in practice but we consider here the exact maximization of the

model evidence. Appendix B describes an algorithm which can be used for this task. The model evidence need not be convex with respect to the different hyper-parameters. Hence, the solution obtained depends on the starting configuration of the optimization algorithm. One typically starts from all hyper-parameters being equal to infinity and iteratively optimizes each hyper-parameter until convergence.

The group ARD formulation (more precisely the type-II maximum likelihood principle followed by group ARD) can be seen as a Bayesian approach where the posterior distribution of each hyper-parameter α_g, with $j = 1, \ldots, d$, is approximated by a delta function centered at the peak of the exact posterior under the assumption of a flat prior for each α_j (Bishop, 2006). Thus, the actual prior for $\mathbf{w}_g$ is a hierarchical mixture where the hyper-prior for α_g is set to be flat (actually flat in log-scale) (Tipping, 2001). If we marginalize out α_g, the actual prior can be shown to become the improper prior:

$$\mathcal{P}(\mathbf{w}_g) \propto \int \mathcal{N}(\mathbf{w}_g|\mathbf{0}, \alpha_g^{-1}\mathbf{I}) d\alpha_g/\alpha_g \propto 1/||\mathbf{w}_g||_2^{d_g} .$$

where $1/\alpha_g$ is the flat hyper-prior for α_g in log-scale and d_g is the dimension of $\mathbf{w}_g$. This improper prior favors solutions with all the components of $\mathbf{w}_g$ set equal to zero since it has an infinitely tall spike at the origin. It also promotes solutions with coefficient values far from zero since it has heavy tails. Thus, it enjoys similar properties to those of the group horseshoe.

The posterior distribution of $\mathbf{w}$ under the group ARD model is Gaussian since both the likelihood and the prior are Gaussian. Thus, sequential experimental design can be carried out very easily in this model using techniques similar to those described in Section 4. Finally, note that the group ARD lacks a hyper-parameter to specify the desired level of sparsity at the group level. The uniform prior assumed for α_g can be considered to be optimal when there is no information about the level of sparsity associated to the learning problem. Nevertheless, this prior can be sub-optimal when such information is available beforehand or when it can be estimated from the data, for example, by cross-validation.

5.5 Other Related Methods

Instead of the EP algorithm, it is also possible to use Markov chain Monte Carlo techniques to approximate the posterior distribution of the Bayesian model introduced in Section 2. For this, we only have to interpret the prior described in (3) for each group of model coefficients $\mathbf{w}_g$ as a mixture of two multivariate Gaussians. A first multivariate Gaussian with zero variance for the different components of $\mathbf{w}_g$ (the spike) and a second multivariate Gaussian with v_0 variance. These two variances are equivalent to the latent parameters λ_g described in Sections 5.2 and 5.3 for the Bayesian group LASSO and the group horseshoe. Thus, we can rely on a Gibbs sampling algorithm very similar to the one described in those sections. Appendix A further details this algorithm. It is inspired from other works for approximate inference in a Bayesian model based on the standard spike-and-slab prior (George and McCulloch, 1997; Lee et al., 2003). Gibbs sampling has also been used by Scheipl et al. (2012) to carry out posterior inference on additive regression models using a generalized prior similar to the one described in this document. Section 7 shows however that EP provides equal performance at a much smaller computational cost. In particular, EP is hundreds of times faster than Gibbs sampling, which is too slow to carry out sequential experimental design.

Variational Bayes (VB) (Attias, 2000; Jaakkola, 2001) is a common alternative for approximate inference in Bayesian models. VB consists in fitting a parametric distribution Q with the aim to ap-

proximate the posterior distribution $\mathcal{P}$. For this, the Kullback-Leibler divergence between Q and $\mathcal{P}$, $\mathrm{KL}(Q||\mathcal{P})$, is minimized. Such an approach differs from EP where the approximate minimization of $\mathrm{KL}(\mathcal{P}||Q)$ takes place. As a result, one can expect VB to be less accurate than EP for the model described in Section 2. Specifically, in sparse linear regression models optimal predictive performance in terms of mean square error is expected to be given by the mean of the posterior distribution. In general, EP produces a global fit to the posterior distribution while VB only approximates the posterior locally around one of its modes. This is illustrated in Bishop (2006). Posterior distributions generated by spike-and-slab priors are often multi-modal and hence the global fit produced by EP is expected to be better at approximating the posterior mean than the local approximation generated by VB. Furthermore, several works in the literature also report a preference of EP over VB in terms of accuracy of the posterior approximation (Minka, 2001; Nickisch and Rasmussen, 2008).

A different generalization of the standard spike-and-slab prior has been proposed in a multi-task setting by Hernández-Lobato et al. (2010). They describe a Bayesian model for the selection of features that are relevant for prediction across L classification tasks. These tasks share the same d features, although the feature values might be different across tasks. The relevant features are identified by using standard spike-and-slab priors on the coefficients of each task, where these priors share the same binary latent variables $\mathbf{z} = (z_1, \ldots, z_d)^{\mathrm{T}}$. These latent variables indicate whether the corresponding features are used for classification in all the tasks or in none of them. This prior cannot be used to tackle the group feature selection problem considered here. In particular, it does not allow to introduce prior knowledge on groups of features that are believed to be jointly relevant or irrelevant for the same task. By contrast, the prior considered in this work can be used to address the multi-task feature selection problem by reformulating it as a single-task learning problem on an extended space:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathbf{X}_L \end{pmatrix}, \qquad \mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_L \end{pmatrix}, \qquad \mathbf{w} = \begin{pmatrix} \mathbf{w}_1 \\ \vdots \\ \mathbf{w}_L \end{pmatrix},$$

where $\mathbf{X}_l$, $\mathbf{y}_l$ and $\mathbf{w}_l$, with $1 \leq l \leq L$, respectively denote the training instances, the targets and the model coefficients for the l-th learning task. The equivalence holds provided that there is a group per feature and that the g-th group contains the L coefficients associated with the g-th feature, for $g = 1, \ldots, d$.

The generalized spike-and-slab prior described in Section 2 can be seen as a degenerate case of the network-based prior proposed by Hernández-Lobato et al. (2011). These authors suggest a modification of the standard spike-and-slab prior by introducing prior dependencies among the components of $\mathbf{z} = (z_1, \ldots, z_d)^{\mathrm{T}}$, that is, the vector of latent variables which indicate whether to use or not each feature for prediction. These dependencies are determined by a network of features. Whenever two features are connected in this network, the two corresponding latent variables are positively correlated. The amount of correlation is specified by a positive parameter b. When the network of features contains G connected components, each component being composed of the features of the same group, and b tends to $+\infty$, the network-based prior is equivalent to the prior considered in this work. However, the network-based prior suggested by Hernández-Lobato et al. (2011) does not provide a direct estimate of the relevance of each group. It only computes posterior probabilities for individual features. Furthermore, the approximate inference mechanism used does not take into account correlations among the different components of $\mathbf{w}$. Specifically, the approximation considered for the posterior of $\mathbf{w}$ factorizes among the different components

of this vector. In the model described in this document these correlations are proven to be very useful to perform sequential experimental design, that is, to determine which instance to include in the training set to obtain the most information about $\mathbf{w}$ (see Section 7). Additionally, the model considered by Hernández-Lobato et al. (2011) can only address binary classification problems and was not designed to address regression problems, which is the focus of the present document.

The work of Yen and Yen (2011) describes an alternative generalization of the spike-and-slab prior which considers both sparsity at the group and the feature level. Specifically, these authors introduce two sets of latent variables. A first set is used to describe whether or not each group of variables is used for prediction, and a second set is used to describe whether or not each feature within a group is used for prediction. Thus, the prior considered in this work can be seen as a particular case of the prior considered by these authors where there is no sparsity at the feature level, but only at the group level. To infer the model coefficients from the data Yen and Yen (2011) do not follow a complete Bayesian approach, and instead find the *maximum a posteriori* (MAP) solution using a blockwise coordinate ascent algorithm. Finding the MAP solution in such model is arguably controversial. In particular, the posterior distribution includes delta functions which take infinite values at some positions and make the objective unbounded. Furthermore, this task involves solving a combinatorial problem which is NP-Hard. To address this difficulty Yen and Yen (2011) propose to use a majorization-minimization technique to simplify the computations needed. Finally, even though they can approximate the MAP solution, their approach does not provide an estimate of the correlations among the different components of $\mathbf{w}$ which are required to carry out sequential experimental design, as described in Section 4.

6. Analysis of Group Sparsity

In this section we study the properties of the generalized spike-and-slab prior to favor solutions that are sparse at the group level. The alternative priors described in Section 5, which can also be used for this purpose, are also analyzed in detail for the sake of comparison. The analysis introduced is based on the work of Carvalho et al. (2009) about the sparsity properties of the standard horseshoe prior and shows interesting insights about the regularization process enforced by each prior distribution and the potential benefits and drawbacks for group feature selection.

Consider the vector $\mathbf{w}_g$ summarizing the d_g model coefficients corresponding to the features contained in the g-th group. The different priors for $\mathbf{w}_g$ analyzed in this section are displayed in Table 1 alongside with their associated hyper-parameters. In this table $\mathcal{N}(\cdot|\boldsymbol{\mu},\boldsymbol{\Sigma})$ denotes a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$; $\mathcal{C}^{+}(\cdot|a,b)$ denotes a half-Cauchy distribution with location and scale parameters a and b, respectively; $\delta(\cdot)$ denotes a point probability mass evaluated at the origin; and $||\cdot||_2$ denotes the ℓ_2-norm. We do not include the prior for the group LASSO since it is identical to the one in the Bayesian group LASSO, as described in Section 5.2. Figure 1 shows the different priors displayed in Table 1 for some values of their hyper-parameters and for a group of size two, that is, $d_g = 2$. In this figure an arrow denotes a point probability mass at the origin. Note that most priors have an infinitely tall spike at the origin to favor solutions with all the model coefficients near or equal to zero. The only exception is the prior corresponding to the group LASSO, which has a sharp peak at the origin instead. From these priors, only the spike-and-slab is able to put a positive probability mass at the origin. This probability is specified by the hyper-parameter $p_{0,g}$. By contrast, in the group horseshoe, the group ARD and the Bayesian group LASSO the probability of observing $\mathbf{w}_g$ at the origin is zero. This means that one will never observe

actual zeros in the samples from these priors. On the other hand, an appealing property of the group ARD and the group horseshoe is that they allow for values of $\mathbf{w}_g$ located far from the origin since they have heavy tails. This is not the case of the spike-and-slab. However, the hyper-parameter that models the variance of the slab, v_0, can be made arbitrary large to account for coefficients significantly different from zero without changing the desired level of sparsity, specified by $p_{0,g}$. The group ARD, the spike-and-slab and the group horseshoe are hence expected to be effective for inducing sparsity at the group level. Specifically, for some value of the prior hyper-parameters, they will either strongly drive the values of the model coefficients towards the origin (as a consequence of the spike), or they will leave them barely unchanged (as a consequence of the heavy tails or the large variance of the slab). The prior corresponding to the Bayesian group LASSO has neither heavy tails nor an infinitely tall spike at the origin. Thus, this prior is expected to perform worse in this task. Finally, the ARD prior is not fully adequate to our problem as it does not include a hyper-parameter to set the desired level of group sparsity, which can be strongly problem dependent.

Prior for $\mathbf{w}_g$	**Density**	**Hyper-parameters**
Generalized spike-and-slab	$p_{0,g}\mathcal{N}(\mathbf{w}_g\|\mathbf{0}, v_0\mathbf{I}) + (1-p_{0,g})\delta(\mathbf{w}_g)$	v_0, $p_{0,g}$
Group horseshoe	$\int \mathcal{N}(\mathbf{w}_g\|\mathbf{0}, \lambda_g^2\tau^2\mathbf{I})\mathcal{C}^+(\lambda_g\|0,1)d\lambda_g$	τ
Bayesian group LASSO	$\propto \left(d_g\gamma^2\right)^{\frac{d_g}{2}} \exp\left\{-\gamma\sqrt{d_g}\|\|\mathbf{w}_g\|\|_2\right\}$	γ
Group ARD	$\propto 1/\|\|\mathbf{w}_g\|\|_2^{d_g}$	-

Table 1: Description of the different priors for enforcing sparsity at the group level.

6.1 Shrinkage Interpretation of the Prior Distributions

The different prior distributions for $\mathbf{w}_g$ that are displayed in Table 1 can be understood as a scale mixture of multivariate Gaussian distributions. More precisely, these priors are equivalent to a zero-mean multivariate Gaussian with a random covariance matrix $\lambda_g^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. The prior distribution for λ_g^2 determines the resulting family of prior distributions for $\mathbf{w}_g$. Thus, under this representation, we have to marginalize out λ_g^2 to evaluate the actual prior probability density for $\mathbf{w}_g$. In particular,

$$\mathcal{P}(\mathbf{w}_g) = \int \mathcal{N}(\mathbf{w}_g|\mathbf{0}, \lambda_g^2\mathbf{I})\mathcal{P}(\lambda_g^2)d\lambda_g^2\,,$$

where $\mathcal{P}(\lambda_g^2)$ denotes the specific prior distribution for λ_g^2. The shrinkage properties of each of the prior distribution displayed in Table 1 can be analyzed by looking at the corresponding assumed prior density for λ_g^2. This density is displayed in Table 2 for each different prior for $\mathbf{w}_g$.

For simplicity, we focus in this section on a toy regression problem which can be analyzed in detail. This problem gives interesting insights about the shrinkage properties of each prior distribution for $\mathbf{w}_g$. In particular, we assume that there is a single group of d_g model coefficients, that is, $\mathbf{w} = \mathbf{w}_g$. Furthermore, we assume that there are $n = d_g$ observations $\mathbf{y}^{\mathrm{T}} = (y_1, \ldots, y_n)$, one for each model coefficient, which are generated according to the rule described in (1) for $\sigma_0^2 = 1$. We also assume that the design matrix $\mathbf{X}$ is equal to the identity matrix $\mathbf{I}$. Under these settings, the optimal value for $\mathbf{w}_g$ is $\mathbf{y}$. Moreover, the expected posterior value for $\mathbf{w}_g$ can be computed exactly given λ_g^2.

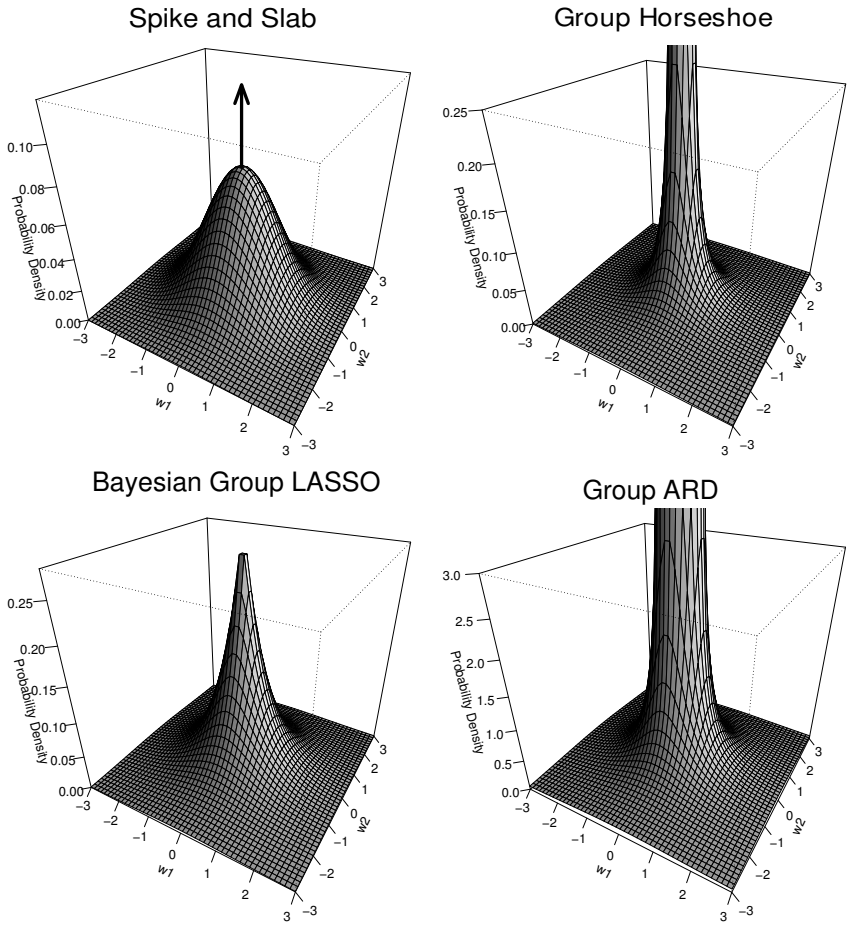

Figure 1: Plots of the different priors displayed in Table 1 for favoring solutions which are sparse at the group level. Results are displayed for $\mathbf{w}_g = (w_1, w_2)^{\mathrm{T}}$, that is, a group of size two, and for some particular values of the hyper-parameters of each different prior distribution. The arrow indicates a point probability mass at the origin. All priors except the prior for the group LASSO and the Bayesian group LASSO have an infinitely tall spike at the origin. The different hyper-parameters are set as follows: $p_{0,g} = 0.5$, $v_0 = 1$, $\tau = 1$ and $\gamma = 1$.

Namely,

$$\mathbb{E}[\mathbf{w}_g|\lambda_g^2] = \frac{\lambda_g^2}{1+\lambda_g^2}\mathbf{y} + \frac{1}{1+\lambda_g^2}\mathbf{0} = \frac{\lambda_g^2}{1+\lambda_g^2}\mathbf{y}, \tag{36}$$

where $\kappa = 1/(1+\lambda_g^2)$, with $\kappa \in [0,1]$, is a random shrinkage coefficient which can be understood as the amount of weight that the posterior mean places at the origin once the targets $\mathbf{y}$ are observed (Carvalho et al., 2009). If $\kappa = 1$ the posterior mean is completely shrunk towards the origin. If $\kappa = 0$, the posterior mean is not regularized at all. Since κ is a random variable, it is possible to plot its prior density to analyze the shrinkage properties of each prior. This density is fully specified

Prior for $\mathbf{w}_g$	**Prior density for λ_g^2**
Generalized spike-and-slab	$p_{0,g}\delta(\lambda_g^2 - v_0) + (1 - p_{0,g})\delta(\lambda_g^2)$
Group horseshoe	$\mathcal{C}^+\left(\sqrt{\lambda_g^2}|0,\tau\right) 1/\left(2\sqrt{\lambda_g^2}\right)$
Bayesian group LASSO	$\text{Gamma}\left(\lambda_g^2 \mid \frac{d_g+1}{2}, \frac{2}{d_g\gamma^2}\right)$
Group ARD	$\propto 1/\lambda_g^2$

Table 2: Description of the different priors assumed for λ_g^2.

by the prior distribution for λ_g^2 which can be any of the ones displayed in Table 2, depending on the actual prior for $\mathbf{w}_g$. In an ideal situation, $\mathcal{P}(\kappa)$, that is, the prior distribution for κ, should favor the bi-separation of the model coefficients that is characteristic of sparse models at the group level. Specifically, while most groups of model coefficients take values close to zero, a few of them take values significantly different from zero. Thus, $\mathcal{P}(\kappa)$ should be large for values of κ near one, to favor the shrinkage of un-important groups of model coefficients. Similarly, $\mathcal{P}(\kappa)$ should be large for values of κ near zero, to barely shrink those groups of model coefficients which are important for prediction.

Figure 2 displays for each different prior distribution for $\mathbf{w}_g$, the corresponding prior distribution for the shrinkage coefficient κ, $\mathcal{P}(\kappa)$. The plots are displayed for a single group of size two. However, similar results are obtained for groups of larger sizes. The prior distributions are obtained from the densities displayed in Table 2 by performing a change of variable since $\kappa = 1/(1+\lambda_g^2)$. For each prior distribution, the corresponding hyper-parameters are selected so that the distance between the 10% and 90% percentiles of the resulting marginal distribution of each component of $\mathbf{w}_g$ is equal to 0.7, 3.5 and 17.5, respectively. These values correspond to *high*, *medium*, and *low* sparsity at the group level. The exception is the prior for κ corresponding to group ARD, which does not have any hyper-parameter to specify the desired level of group sparsity. In this figure the arrows denote a point probability mass and the length of the arrow is proportional to the corresponding probability mass.

Figure 2 shows that the prior corresponding to the Bayesian group LASSO is not able to simultaneously produce large densities for values of κ close to zero and one. Furthermore, the probability density for $\kappa = 1$ is always equal to zero. This is an unexpected result which questions the capacity of this prior to provide solutions that are sparse at the group level in a selective manner. In particular, under this prior it is not possible to achieve high levels of sparsity (this corresponds to high density values of κ near one) while not shrinking the model coefficients that are different from zero towards the origin (this corresponds to high density of κ near the origin). The next section illustrates that these issues still appear when the MAP solution is used and zeros are produced in the solution set, as in the group LASSO.

The other priors, that is, the spike-and-slab, the group horseshoe and the group ARD, do not suffer from the limitations described for the Bayesian group LASSO. These priors produce densities that are peaked at $\kappa = 1$ and at values of κ near the origin. Furthermore, in the case of the spike-and-slab prior, one actually obtains a positive probability at $\kappa = 1$. Thus, the posterior distribution of the coefficients corresponding to non-predictive features will concentrate near the origin under these priors. On the other hand, both the group horseshoe and the group ARD priors are characterized by heavy tails. This can be observed in Figure 2 by the fact that they simultaneously

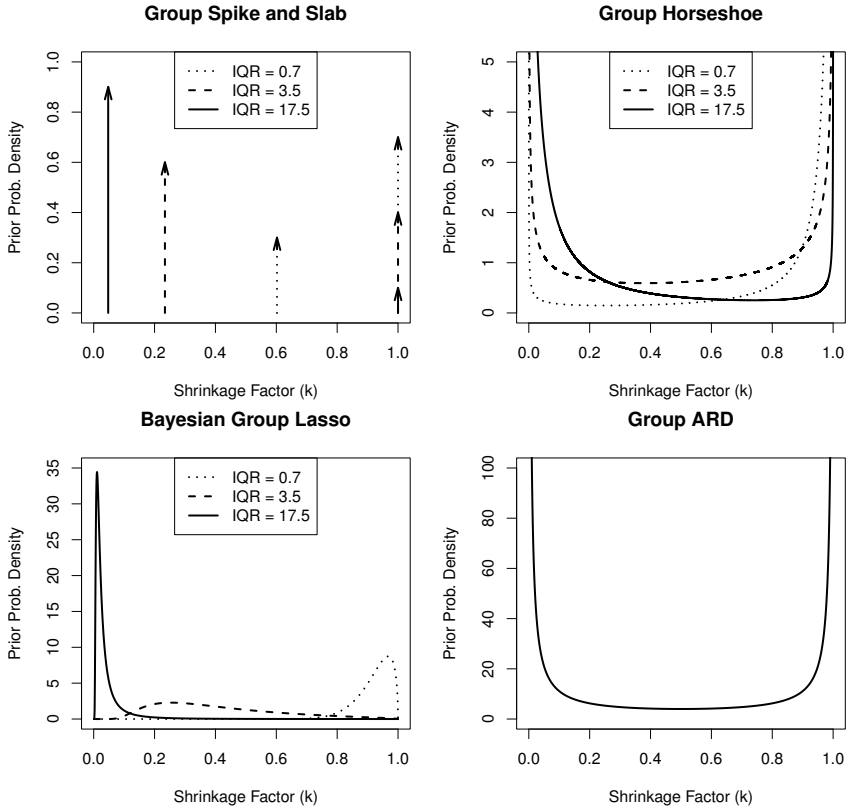

Figure 2: Prior distribution for κ associated to the different priors for $\mathbf{w}_g$ displayed in Table 1. The plots are displayed for a single group of size two, and for some particular values of the hyper-parameters of each prior distribution for $\mathbf{w}_g$ that give the same inter-quantile range (IQR) for each individual component of this vector. A single curve is plotted for the group ARD prior since it does not have any hyper-parameter. An arrow denotes a point probability mass.

give high probabilities to values of κ near the origin. The consequence is that these priors will barely regularize important groups of coefficients that are strictly needed for prediction. The spike-and-slab prior does not have heavy tails. However, a similar effect can be obtained by specifying large values for v_0, the parameter that controls the variance of the slab. Therefore, these three prior distributions are expected to selectively shrink the posterior mean, which is the ideal situation for regression problems which are sparse at the group level.

6.2 Regularization Properties of the Prior Distributions

We continue the analysis of the toy regression problem described in Section 6.1. Specifically, we study the behavior of the posterior mean, $\mathbb{E}[\mathbf{w}_g]$, under the different priors for $\mathbf{w}_g$ when the targets $\mathbf{y}$ are similar to or very different from the prior mean, that is, a vector with all the components equal to zero. It is possible to show, by marginalizing (36) over the posterior distribution for λ_g^2, that in this toy problem $\mathbb{E}[\mathbf{w}_g]$ is a vector parallel to $\mathbf{y}$. More precisely, these two vectors only

differ in their ℓ_2-norms. In particular, the ℓ_2-norm of $\mathbb{E}[\mathbf{w}_g]$ is always smaller or equal than the ℓ_2-norm of $\mathbf{y}$. The same applies when the MAP estimate is considered instead of the posterior mean, as in the group LASSO. Thus, we can analyze the regularization properties of each prior distribution by comparing the ℓ_2-norm of the targets $||\mathbf{y}||_2$ with the corresponding value of the ℓ_2-norm of the posterior mean $||\mathbb{E}[\mathbf{w}_g]||_2$. Figure 3 shows a comparison between these two norms for each different prior distribution for $\mathbf{w}_g$ and for different values of the prior hyper-parameters. The identity function is also displayed to indicate where the two norms are equal. We include plots both for the group LASSO and the Bayesian group LASSO to illustrate the differences between the posterior mean $\mathbb{E}[\mathbf{w}_g]$ and the MAP estimate $\hat{\mathbf{w}}_g$. Similarly, since the spike-and-slab prior has two hyper-parameters, $p_{0,g}$ and v_0, we display two plots for this prior. In the first plot $p_{0,g}$ varies and v_0 is kept constant. Conversely, in the second plot v_0 varies and $p_{0,g}$ is kept constant. Finally, we provide a single curve for the group ARD prior since it does not have any hyper-parameter.

Figure 3 shows that for small values of $||\mathbf{y}||_2$, the group LASSO actually drives the model coefficients to zero. In particular, under this model it is possible to show that if $||\mathbf{y}||_2 \leq \sqrt{d_g}\gamma/2$ then $||\hat{\mathbf{w}}_g||_2$, the ℓ_2-norm of the MAP estimate of $\mathbf{w}$, is equal to zero. When $||\mathbf{y}||_2 > \sqrt{d_g}\gamma/2$, the ℓ_2-norm of $\hat{\mathbf{w}}_g$ must satisfy $||\hat{\mathbf{w}}_g||_2 = ||\mathbf{y}||_2 - \sqrt{d_g}\gamma/2$. Thus, if the targets $\mathbf{y}$ are significantly different from the mean of the prior, the differences between the two norms in the group LASSO are actually constant and proportional to the value of the hyper-parameter γ, which controls the level of sparsity. The consequence is that for high levels of group sparsity, as specified by γ, the group LASSO regularizes the coefficients that are different from zero and introduces a significant bias in their estimation. Specifically, under this model it is not possible to simultaneously consider large values for the model coefficients and high levels of group sparsity. When the posterior mean is considered instead of the MAP estimate, as in the Bayesian group LASSO, the observed behavior is very similar for large values of $||\mathbf{y}||_2$. Nevertheless, in this case, small values of $||\mathbf{y}||_2$ no longer produce zeros in the estimation for $\mathbf{w}$, but a stronger regularization effect which forces $\mathbb{E}[\mathbf{w}_g]$ to be closer to the origin. These two methods, that is, the group LASSO and the Bayesian group LASSO, are hence unable to shrink the model coefficients in a selective manner and are expected to lead to an impaired prediction performance in problems that are actually sparse at the group level.

The group ARD also drives the model coefficients towards zero for small values of $||\mathbf{y}||_2$. Specifically, if $||\mathbf{y}||_2 \leq \sqrt{d_g}$ then the optimal parameter λ_g^2 which maximizes the model evidence is equal to zero, and the posterior estimate of $\mathbf{w}$, $\mathbb{E}[\mathbf{w}_g]$, is placed at the origin. When $||\mathbf{y}||_2 > \sqrt{d_g}$, the optimal value for λ_g^2 is equal to $||\mathbf{y}||_2^2/d_g - 1$ and hence, from (36), $||\mathbb{E}[\mathbf{w}_g]||_2 = ||\mathbf{y}||_2 - d_g/||\mathbf{y}||_2$. Thus, the group ARD also introduces a bias in the estimation of the model coefficients when the targets are significantly different from the origin. Nevertheless, this bias is equal to $d_g/||\mathbf{y}||_2$ and tends to zero when $||\mathbf{y}||_2$ approaches infinity. This is a consequence of the heavy tails of the prior for $\mathbf{w}_g$ which barely regularizes the model coefficients when these are significantly different from zero and strictly required for prediction. A similar behavior is observed for the group horseshoe. Namely, when $||\mathbf{y}||_2$ approaches infinity both norms tend to coincide and the prior distribution barely regularizes the model coefficients. By contrast, for small values of $||\mathbf{y}||_2$, the model coefficients are strongly regularized in an amount that depends on τ. The smaller its value, the stronger the regularization effect. Note that this parameter has very little effect on the regularization of the model coefficients when $||\mathbf{y}||_2$ is large. This is a very interesting property. From this, we conclude that these two prior distributions, the group ARD and the group horseshoe, are expected to be useful to provide the bi-separation of the model coefficients that is characteristic of sparse models at the group level.

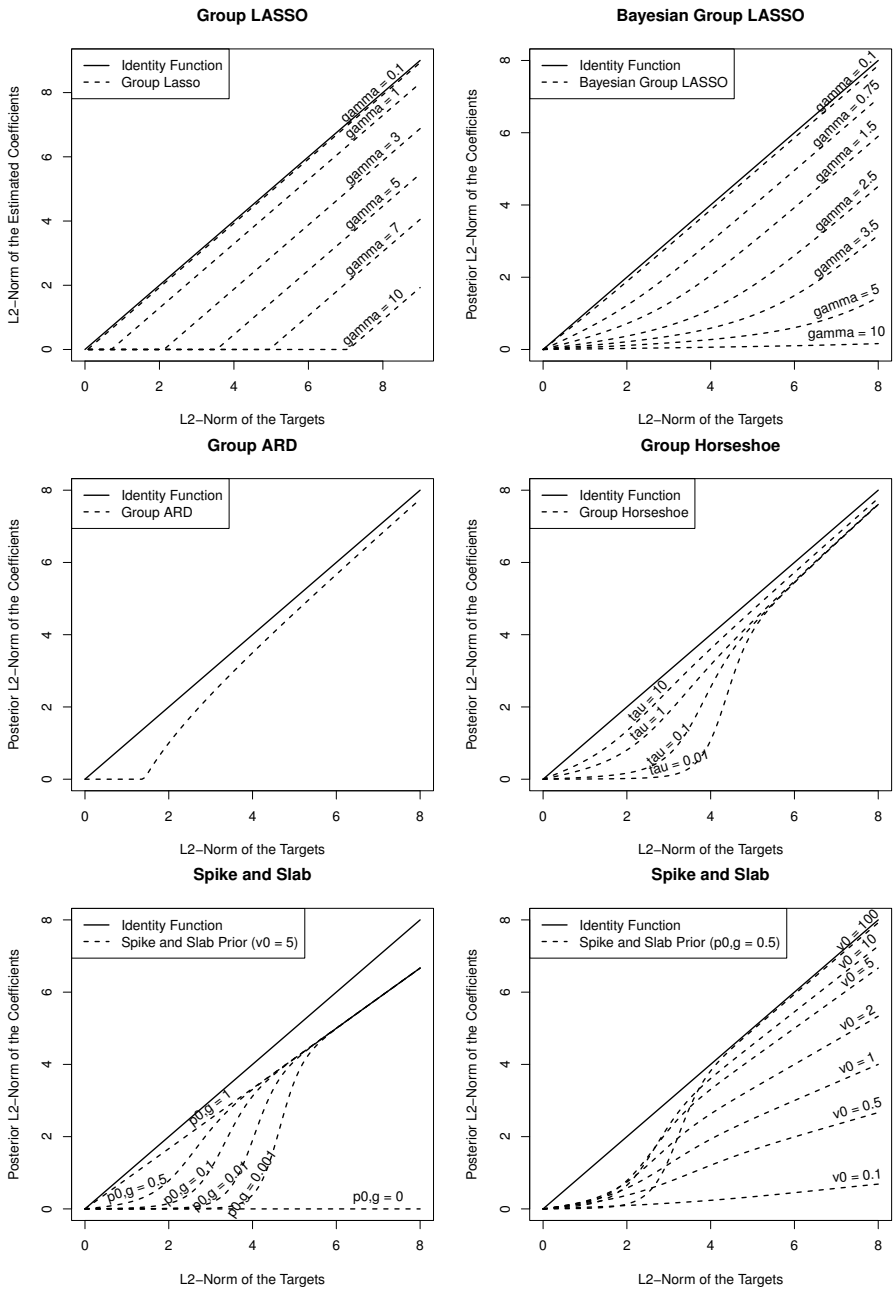

Figure 3: Plots of the ℓ_2-norm of the posterior expected value of the model coefficients, as a function of the ℓ_2-norm of the observed targets $\mathbf{y}$, for each different prior distribution for $\mathbf{w}_g$, and for different values of the hyper-parameters of each prior. We report results for both the Bayesian group LASSO and for the group LASSO, that is, the MAP estimate in the Bayesian group LASSO. The diagonal solid line represents where both norms are equal.

In the spike-and-slab prior, modifying $p_{0,g}$, that is, the parameter that determines the prior probability that the model coefficients are equal to zero, has an impact in the posterior mean when the optimal values of the model coefficients are close to zero. In particular, reducing this hyper-parameter produces a stronger regularization effect for small values of $||\mathbf{y}||_2$ without affecting the regularization of the model coefficients when these are significantly different from zero, or equivalently, when $||\mathbf{y}||_2$ is large. An exception is observed when $p_{0,g}$ is set equal to zero. In this case, the posterior mean is placed at the origin. The modification of v_0, that is, the parameter that determines the variance of the slab, has little impact when $||\mathbf{y}||_2$ is close to zero. By contrast, this parameter fully specifies the regularization of the model coefficients that are strictly important for prediction. Specifically, for large $||\mathbf{y}||_2$ it is possible to show that the posterior distribution for λ_g^2 tends to $\delta(\lambda_g^2 - v_0)$ and consequently, from (36), that $\mathbb{E}[\mathbf{w}_g] \approx \mathbf{y} - \mathbf{y}/(1+v_0)$. This indicates that the spike-and-slab prior introduces a positive bias in the estimation of the model coefficients, when these are far from the origin. This bias is proportional to the optimal value of each coefficient and it is a consequence of the absence of heavy tails in this prior distribution to explain for large values of the model coefficients. Nevertheless, we note that it is possible to reduce the estimation bias simply by increasing v_0, as illustrated by Figure 3. In summary, the spike-and-slab is also expected to perform well in problems that are sparse at the group level. In particular, this prior distribution is also able to model the bi-separation of the model coefficients that is characteristic of this type of problems.

7. Experiments

In this section, the performance of the model based on the EP algorithm and the generalized spike-and-slab prior is evaluated in several regression problems from different domains of application, using both simulated and real-world data. The problems analyzed include the reconstruction of sparse signals from a reduced number of noisy measurements (Huang and Zhang, 2010; Ji et al., 2008), the prediction of user sentiment from customer-written reviews of kitchen appliances and books (Pang et al., 2002; Blitzer et al., 2007) and the reconstruction of images of hand-written digits extracted from the MNIST data set (LeCun et al., 1998). The data sets of these problems have similar characteristics. That is, a large number of attributes and a rather small number of training instances, that is, $d \gg n$. Similarly, on each data set only a reduced number of features is expected to be useful for prediction. These are precisely the characteristics of the regression problems where the sparsity assumption is expected to perform well and to be useful for induction. In these experiments, the prior information about groups of features that are expected to be jointly relevant or jointly irrelevant for prediction is assumed to be given or it is estimated from additional data.

We refer to the regression model that assumes generalized spike-and-slab priors and uses EP for approximate inference as GSS-EP. This model is compared in this section with the related methods for group feature selection described in Section 5. Namely, the group LASSO (G-LASSO), the Bayesian group LASSO (BG-LASSO), the group horseshoe (G-HS), the group ARD formulation (G-ARD), a model that also assumes generalized spike-and-slab priors but uses Markov chain Monte Carlo sampling for approximate inference (GSS-MCMC) and finally, a regression model that also uses EP for approximate inference but only considers the standard spike-and-slab prior for induction (SS-EP). This last model is described in detail by Hernández-Lobato (2010) and is a particular case of GSS-EP which does not consider the grouping information in the induction process,

that is, all groups are of size one. SS-EP is included in the comparison to evaluate the benefit of considering sparsity at the group level instead of only at the feature level. Furthermore, comparing results with respect to SS-EP is also supported by the good performance obtained by such method in regression problems that are sparse at the feature level (Hernández-Lobato, 2010). Similarly, GSS-MCMC is included in the comparison to evaluate the performance of the EP approximation of the posterior distribution. In particular, Markov chain Monte Carlo methods do not suffer from any approximation bias, unlike deterministic techniques, such as the EP algorithm.

In our experiments, we report the training time of each method being evaluated. The different training algorithms have been implemented in R (R Development Core Team, 2011), and care has been taken to make them as efficient as possible.[1] Specifically, the implementation of G-LASSO is based on the fast algorithm described by Roth and Fischer (2008) and Appendix A gives all the details about the implementation of BG-LASSO, G-HS and GSS-MCMC. In these three methods, the posterior distribution of the model is approximated by generating 10,000 Gibbs samples after a burn-in period of 1,000 samples. This number of samples seems to be adequate and experiments indicate that no significant improvements are obtained by increasing the number of samples generated. All the details of the implementation of G-ARD are given in Appendix B. Finally, in GSS-EP, GSS-MCMC, SS-EP, G-ARD, BG-LASSO and G-HS the estimate of the model coefficients, $\hat{\mathbf{w}}$, is given by the approximate posterior mean. In G-LASSO $\hat{\mathbf{w}}$ is given by the MAP estimate.

7.1 Reconstruction of Sparse Signals

A first batch of experiments is carried out to illustrate the potential applications of the generalized spike-and-slab prior in the field of compressive sensing (Donoho, 2006; Candes and Wakin, 2008). The objective in compressive sensing is to reconstruct a sparse signal, generally codified in the model coefficients $\mathbf{w} = (w_1, \ldots, w_d)^{\mathrm{T}}$, from a limited set of linear measurements $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, with $n \ll d$. The measurements $\mathbf{y}$ are obtained after projecting the signal $\mathbf{w}$ onto an $n \times d$ measurement matrix $\mathbf{X}$, that is, $\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}} \sim \mathcal{N}(\mathbf{0}, \sigma_0^2 \mathbf{I})$ is a Gaussian noise. If $\mathbf{w}$ is sparse, it is possible to reconstruct this vector accurately from $\mathbf{y}$ and $\mathbf{X}$ using fewer measurements than the number of degrees of freedom of the signal, which is the limit required to guarantee the reconstruction of arbitrary signals. When $\mathbf{w}$ is not sparse, it can still be estimated using less than d samples provided that it is compressible in some orthogonal basis $\mathbf{B}$, for example, a wavelet basis, such that $\tilde{\mathbf{w}} = \mathbf{B}^{\mathrm{T}}\mathbf{w}$ is sparse or nearly sparse. In this case, the measurement process is performed after projecting the signal onto the columns of $\mathbf{B}$, that is, $\mathbf{y} = \mathbf{X}\mathbf{B}^{\mathrm{T}}\mathbf{w} + \boldsymbol{\epsilon} = \mathbf{X}\tilde{\mathbf{w}} + \boldsymbol{\epsilon}$. Once an estimate of $\tilde{\mathbf{w}}$ is obtained from $\mathbf{y}$ and $\mathbf{X}$, we can approximate $\mathbf{w}$ using $\mathbf{w} = \mathbf{B}\tilde{\mathbf{w}}$.

We evaluate the different methods in a problem similar to the standard benchmark problems used in the field of signal reconstruction (Ji et al., 2008). More precisely, we generate 100 random sparse signals to be reconstructed from noisy measurements where each signal has $d = 512$ random components that are codified using a particular group sparsity pattern. Specifically, the components of each signal (i.e., the model coefficients) are iteratively assigned to $G = 128$ different groups of the same size that contain 4 components. From the 128 groups of components, only 4 randomly chosen groups contain components that are actually different from zero. The values of these components are uniformly chosen in the interval $[-1, 1]$. The resulting signal is stored in the vector $\mathbf{w}_0$ of model coefficients. This vector contains the sparse signal to be reconstructed. Given a particular signal $\mathbf{w}_0$ we then generate a reduced amount of measurements using a design matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathrm{T}}$

1. The R source code for GSS-EP is available at http://arantxa.ii.uam.es/%7edhernan/GSS-EP/.

whose rows are sampled uniformly in the hyper-sphere of radius $\sqrt{d}$. For the reconstruction of each signal a total of $n = 64$ noisy measurements $\mathbf{y} = \{y_1, \ldots, y_n\}$ are used. These are generated as $y_i = \mathbf{w}_0^{\mathrm{T}} \mathbf{x}_i + \varepsilon_i$, for $i = 1, \ldots, n$, where ε_i follows a standard Gaussian distribution.

Given $\mathbf{X}$ and $\mathbf{y}$ we induce $\mathbf{w}_0$ using the different methods evaluated. Let $\hat{\mathbf{w}}$ be the corresponding estimate of the signal $\mathbf{w}_0$. The reconstruction error is quantified by $||\hat{\mathbf{w}} - \mathbf{w}_0||_2 / ||\mathbf{w}_0||_2$, where $||\cdot||_2$ denotes the ℓ_2-norm. In these experiments, the hyper-parameters of each method are fixed optimally. In particular, we set p_0, that is, the fraction of groups initially expected to be relevant for prediction, equal to $4/128$ in GSS-EP and GSS-MCMC. In these two methods v_0, that is, the variance of the slab, is set equal to $1/3$, that is, the actual variance of the components of $\mathbf{w}_0$ that are different from zero. In SS-EP the same value is used for v_0, but $p_0 = 16/512$ since in this model there is a group for each different coefficient. In BG-LASSO we set $\gamma = \sqrt{120}$ so that the resulting prior has the same variance as the expected variance of the signal $\mathbf{w}_0$. In G-LASSO we try different values for k and report the best performing value observed, which corresponds to $k = 8$. The group horseshoe prior does not have defined variances. Thus, we fix τ in G-HS so that the marginals under the group horseshoe have the same distance between the percentiles 1% and 99% as the marginals under the generalized spike-and-slab prior. We specifically use these extreme values for the percentiles because otherwise, for high levels of sparsity, that is, when p_0 is close to zero, the spike-and-slab prior can easily have an inter quantile range equal to zero. Finally, in all methods except G-LASSO, σ_0^2, that is, the variance of the Gaussian noise, is set equal to one.

The results of these experiments are displayed in Table 3. This table shows the average reconstruction error of each method and the corresponding average training time in seconds.[2] The figures after the symbol $\pm$ are standard deviation estimates. We note that GSS-EP obtains the best reconstruction error. Furthermore, the performance of this method is equivalent to the performance of GSS-MCMC. This indicates that the posterior approximation obtained by EP is accurate. The table also illustrates the importance of considering the grouping information for prediction. In particular, SS-EP obtains a significantly worse reconstruction error. After GSS-EP and GSS-MCMC, the best performing methods are GS-HS and G-ARD. The reconstruction error of these methods is only slightly worse than the reconstruction error obtained when generalized spike-and-slab priors are assumed. Finally, we note that the performance of G-LASSO, and especially BG-LASSO, is significantly worse than the performance of the other methods that use the grouping information. This validates the results of Section 6, where these two methods were expected to perform poorly. A paired Student t-test confirms that GSS-EP and GSS-MCMC perform better than the other methods being compared (p-value below 5%). On the contrary, the differences in performance between these two methods are not statistically significant.

	GSS-EP	GSS-MCMC	SS-EP	G-LASSO	BG-LASSO	G-HS	G-ARD
Error	0.29±0.11	0.29±0.10	0.71±0.20	0.54±0.11	0.92±0.03	0.35±0.12	0.39±0.12
Time	1.60±1.91	1168±64	3.89±3.42	2.39±1.63	3007±339	2909±328	2.56±1.55

Table 3: Average reconstruction error and training time of each method on the sparse signal reconstruction problem.

2. Training times were measured on an Intel(R) Xeon(R) 2.5Ghz CPU.

When comparing the average training time of the different methods displayed in Table 3 we observe that the fastest method is GSS-EP. In particular, GSS-EP needs approximately one second and a half for training, on average. This method is more than 500 times faster than GSS-MCMC and more than 1,000 times faster than BG-LASSO or G-HS, the induction methods that rely on Gibbs sampling for approximate inference. The training time of G-LASSO and G-ARD is also small, typically below 10 seconds, but above the training time of GSS-EP. Finally, we note that the training time of SS-EP exceeds the training time of GSS-EP. This is because in SS-EP the EP algorithm requires more iterations to converge, even though SS-EP is implemented with the same code as GSS-EP and also uses damped EP updates to improve convergence. This specific problem is also reported by Hernández-Lobato (2010) and is a consequence of the multiple modes of the posterior distribution under the standard spike-and-slab prior. When the grouping information is considered for induction, as in GSS-EP, the problem seems to be alleviated and the EP algorithm converges in fewer iterations.

Figure 4 displays, for a given instance of the signal reconstruction problem, the actual signal $\mathbf{w}_0$ and the different reconstructions generated by each method. The figure shows that GSS-EP and GSS-MCMC obtain the most accurate reconstructions and only fail to reconstruct the smallest components of the signal, probably due to the measurement noise. Furthermore, the reconstructions from these two methods look nearly identical. This gives further evidence indicating that the EP posterior approximation of the posterior mean is accurate. By contrast, SS-EP completely fails to reconstruct the original signal. In particular, without the grouping information it is impossible to reconstruct accurately this signal and SS-EP actually includes in the solution coefficients that only explain the observed data by chance. Similarly, many important coefficients are excluded from the solution found by SS-EP. The reconstructions generated by G-HS and G-ARD look accurate too. However, these methods include many coefficients with values just slightly different from zero which were not present in the original signal. In the group horseshoe this behavior can be explained because under this prior the probability of observing a group at the origin is zero. Similarly, in G-ARD the optimization process is not convex and can converge to a local and sub-optimal maximum of the type-II likelihood. The signal reconstructed by G-LASSO underestimates some values of the model coefficients and, at the same time, includes many coefficients that take values slightly different from zero. This behavior is due to the properties of the corresponding Multi-variate Laplace prior described in Section 6. In particular, this prior distribution is unable to achieve high levels of sparsity without shrinking too much the model coefficients that are different from zero. Finally, the signal reconstructed by BG-LASSO is rather inaccurate and not sparse at all. The Multi-variate Laplace prior produces an excessive shrinkage of non-zero model coefficients, while the magnitude of the coefficients that should be zero is not sufficiently reduced. Specifically, when a full Bayesian approach is used under this prior, the probability density of observing a group at the origin is zero, as illustrated by Figure 2. This questions the utility of the Bayesian group LASSO for group feature selection.

We also evaluate the utility of sequential experimental design to generate the different measurements used for the reconstruction of the sparse signal. For this, we repeat the previous experiments for an iteratively increasing number of measurements and report the reconstruction error of each method, except BG-LASSO, G-HS and GSS-MCMC. These methods are excluded because the cost of Gibbs sampling makes the corresponding computations too expensive. In these experiments, we start from an initial set of 32 measurements which are randomly generated using a design matrix $\mathbf{X}$ whose rows are sampled uniformly in the hyper-sphere of radius $\sqrt{d}$. The reason for this is that the

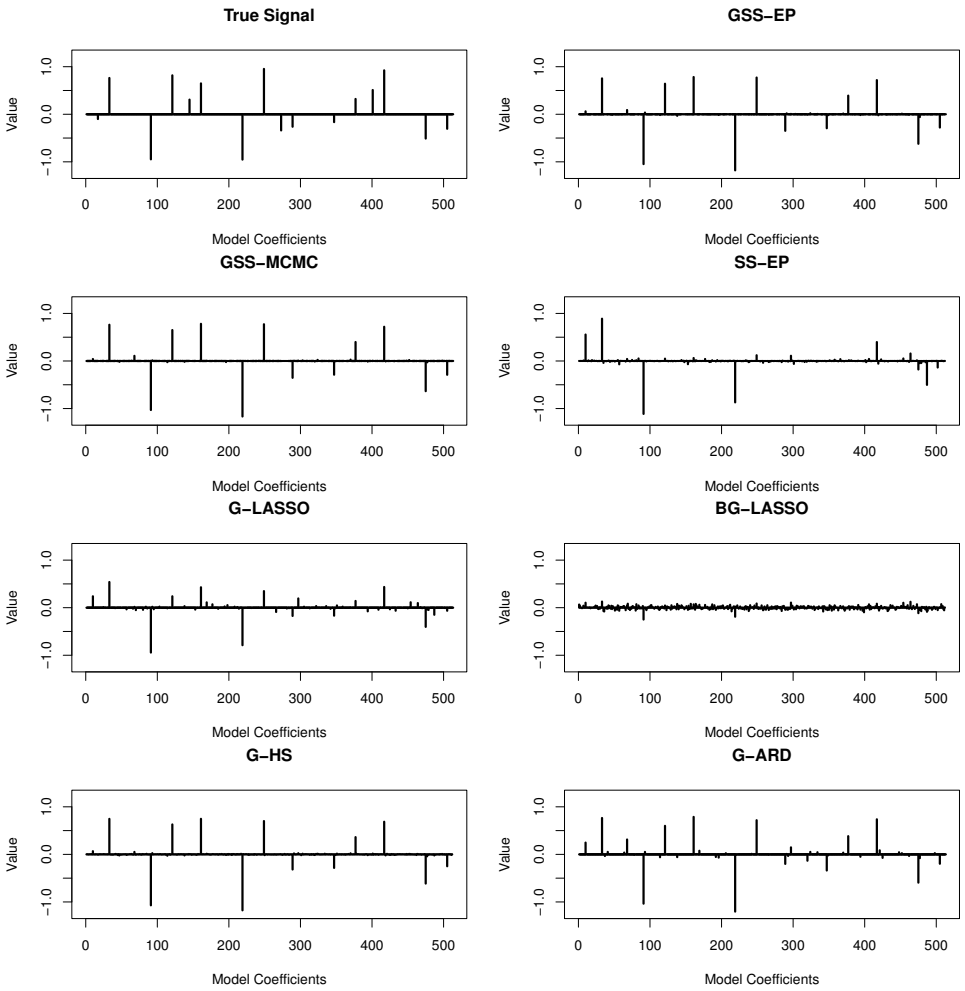

Figure 4: True signal $\mathbf{w}_0$ and the different reconstructions generated by each method for one instance of the signal reconstruction problem. Similar results are observed in other instances of this problem. The x-axis represents each of the model coefficients and the y-axis their values.

sequential design of new measurements from the beginning typically leads to over-fitting, as indicated by Ji and Carin (2007). We then iteratively generate up to 32 extra measurements and report the corresponding reconstruction error. The new measurements are generated using the sequential experimental design strategy described in Section 4. To investigate the benefits of such strategy, we also report results when these new 32 measurements are randomly generated. G-LASSO is also excluded from the experiments since it is impossible to compute posterior covariances under this method.

The results of the experiments are displayed graphically in Figure 5. This figure shows the reconstruction error of GSS-EP, SS-EP, and G-ARD as a function of the number of measurements performed when $\mathbf{X}$ is designed or randomly chosen. These curves indicate that sequential experimental design significantly improves the reconstruction error when compared to random design. In

particular, a steeper error descent with respect to the number of measurements made is produced for GSS-EP, SS-EP and G-ARD. The smallest final reconstruction error is achieved by GSS-EP followed by G-ARD, when $\mathbf{X}$ is designed. Furthermore, the reconstruction error of SS-EP significantly improves in this situation and becomes very close to the reconstruction error of the methods that consider the grouping information when $\mathbf{X}$ is chosen randomly. This illustrates the benefits of sequential experimental design. In any case, the reconstruction error of GSS-EP is much better than the reconstruction error of SS-EP when $\mathbf{X}$ is designed. This remarks again the beneficial properties of considering the grouping information during the induction process.

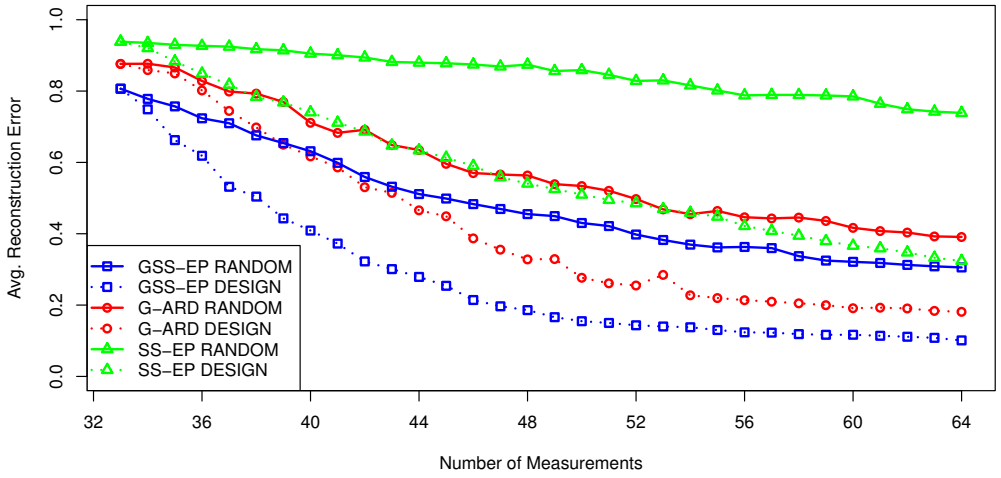

Figure 5: Average reconstruction error on the signal reconstruction problem as a function of the number of measurements performed for GSS-EP, G-ARD and SS-EP. We report results when $\mathbf{X}$ is chosen randomly and when $\mathbf{X}$ is designed.

The last experiments of this section investigate the sensitivity of the generalized spike-and-slab prior to the incorrect specification of the information about which groups of model coefficients are expected to be jointly relevant or jointly irrelevant for prediction. For this, we repeat the experiments whose results are displayed in Table 3. However, this time we introduce increasing levels of noise in the grouping information. Specifically, we permute at random 10%, 20%, 40%, 60% and 100% of the components of a vector with length equal to d, the total number of components in the signal to be reconstructed. Such vector summarizes the grouping information. Namely, its entries take values between 1 and the total number of groups, 128, and they respectively indicate the particular group each component belongs to. When 100% of the components of the vector are randomly permuted, the grouping information is completely random. By contrast, when a smaller fraction of the components are permuted, the vector may still contain useful information that is only partially correct.

Table 4 displays the average reconstruction error of each method for each different level of noise introduced in the grouping information. The error of a method has been high-lighted in bold-face when it is better than the error of GSS-EP and there is statistical evidence indicating a performance

difference. A paired t-test has been used for this purpose (p-value $< 5\%$). Similarly, when there is not enough statistical evidence to indicate a performance difference (GSS-EP performs similarly), the corresponding error has been underlined. The errors where GSS-EP is found to perform better according to the t-test have been left un-modified. The table shows that when the level of noise in the groups is small (i.e., 10% or 20%) the grouping information is still useful to reconstruct the sparse signal and GSS-MCMC and GSS-EP perform best. By contrast, when the level of noise introduced is large (i.e., 60% or 100%), using the grouping information for induction is harmful and the performance of the different methods degrades significantly. The best method is in this case SS-EP as a logical consequence of not considering the grouping information. Similar results have been obtained by Huang and Zhang (2010) for the group LASSO.

Noise Level	GSS-EP	GSS-MCMC	SS-EP	G-LASSO	BG-LASSO	G-HS	G-ARD
10%	0.39±0.17	0.38±0.16	0.71±0.20	0.60±0.13	0.92±0.03	0.44±0.16	0.48±0.16
20%	0.51±0.19	**0.49±0.17**	0.71±0.20	0.65±0.12	0.93±0.02	0.54±0.15	0.56±0.18
40%	0.70±0.24	0.69±0.20	0.71±0.20	0.75±0.11	0.93±0.02	0.68±0.16	0.73±0.19
60%	0.83±0.18	0.83±0.17	**0.71±0.20**	0.81±0.09	0.94±0.02	**0.78±0.12**	0.88±0.17
100%	0.95±0.17	**0.90±0.14**	**0.71±0.20**	**0.85±0.07**	0.94±0.02	**0.86±0.11**	0.97±0.15

Table 4: Average reconstruction error of each method on the sparse signal reconstruction problem as a function of the fraction of components (noise level) that are randomly permuted in the vector that contains the grouping information.

7.2 Prediction of User Sentiment

In this section we investigate the utility of the generalized spike-and-slab prior to address the problem of sentiment prediction from user-written product reviews (Pang et al., 2002). In this task the objective is to predict the rating assigned by a user to a particular product in terms of the text contained in the product review by the user. We specifically focus on the four sentiment data sets described by Blitzer et al. (2007). These data sets contain reviews and the corresponding user ratings from different products extracted from *www.amazon.com*. The products from each data set are classified in a different category: *books*, *DVDs*, *kitchen appliances* or *electronics*, and the range of possible user ratings goes from 1 to 5. Each review is represented by a vector of features whose components correspond to the unigrams and bigrams (i.e., single words and pairs of words, respectively) that appear in at least 100 reviews of the products within the same category. The feature values are simply the number of times that the corresponding unigram or bigram appears in the review. Table 5 displays the total number of instances and the total number of features of each sentiment data set.

The predictive performance of the different methods compared is evaluated on the sentiment data sets *Books* and *Kitchen*. The data sets *DVDs* and *Electronics* are respectively used to generate the different groups of features considered for induction in the latter data sets. In particular, the *DVDs* data set is used to generate the groups for the *Books* data set and the *Electronics* data set is used to generate the groups for the *Kitchen* data set. All the features of the *Book* data set are contained in the *DVDs* data set and all the features of the *Kitchen* data set are contained in the

Data Set	# Instances	# Features (d)
Books	5501	1213
DVDs	5518	1303
Kitchen	5149	824
Electronics	5901	1129

Table 5: Number of instances and features corresponding to each sentiment data set.

Electronics data set. This means that we can safely consider exclusively the common features between these pairs of data sets to generate the groups. For this, we assume that the relevance or irrelevance of each group of features transfers from one data set to another. We then fit a simple linear ridge regression model using all the available data of the data set from which the groups are induced and a hierarchical clustering algorithm is run on the absolute values of the estimated model coefficients. This algorithm is stopped when 150 clusters are generated and the features associated to the coefficients contained in each cluster are grouped together. This guarantees that the features for which the associated model coefficients take similar values are contained in the same group. Thus, we expect that sets of relevant features (i.e., the features whose associated coefficients take large values) are actually placed in the same group of features. The same behavior is expected for irrelevant features (i.e., the features whose associated coefficients take small values). The results of the experiments reported in this section seem to confirm our expectation. In particular, using the grouping information for prediction has a beneficial effect on the predictive performance.

The evaluation procedure on each data set consists in generating 100 random partitions of the data into a training set with $n = 100$ instances and a test set with the remaining data. This particular size for the training set is chosen because we are interested in evaluating the performance of the different methods when $n \ll d$. In each train and test partition the features are normalized alongside with the targets to have zero mean and unit standard deviation across data instances. Furthermore, the hyper-parameters of GSS-EP, p_0 and v_0, are chosen in terms of an independent 10-fold cross-validation estimate of the prediction performance computed on the training data. The model evidence, as estimated by the EP algorithm, is not used for this purpose since it has been empirically found to provide inaccurate decisions in these data sets. The values of p_0 and v_0 are selected from a grid of 5×5 points. This grid is centered on a combination of hyper-parameters with good estimated predictive performances for GSS-EP. In SS-EP p_0 and v_0 are also selected by 10-fold cross-validation, but using a different grid of 5×5 points. This grid is also centered on a combination of hyper-parameters with good estimated predictive performances for SS-EP. In GSS-MCMC we use the same hyper-parameters as the ones found in GSS-EP. This allows to directly compare results between GSS-MCMC and GSS-EP. G-HS and BG-LASSO are too computationally expensive for a cross-validation search of the prior hyper-parameters. In BG-LASSO we set γ so that the marginals of the resulting prior have, on average, the same variances as the variances of the marginals of the generalized spike-and-slab prior in GSS-EP. In G-HS we select τ so that the marginals of the group horseshoe prior have the same distance between the percentiles 1% and 99% as the marginals under the generalized spike-and-slab prior in GSS-EP. Recall that the horseshoe prior does not have defined variances. In G-LASSO we select k by 10-fold cross-validation using a grid of 10 values. This grid contains hyper-parameter values with good estimated predictive performances for G-LASSO. Finally, σ_0^2, that is, the variance of the noise, is set equal to one in all the

methods compared. This specific value provides good results in general. Recall that the targets $\mathbf{y}$ are normalized to have zero mean and unit standard deviation in each train and test partition.

The results of these experiments are displayed in Tables 6 and 7 for the *Books* and the *Kitchen* data sets, respectively. These tables show the mean squared error (MSE) of each method on the test set and the corresponding average training time in seconds, without including the time required for finding the model hyper-parameters. The tables show that in both data sets GSS-EP obtains nearly the best prediction results. Furthermore, the prediction error of GSS-EP is again almost equivalent to the prediction error of GSS-MCMC. This gives further evidence supporting that the posterior approximation computed by EP is accurate. These results also indicate that it is important to consider the grouping information for prediction. In particular, SS-EP obtains worse reconstruction errors than GSS-EP in both data sets. After GSS-EP and GSS-MCMC, the best performing method is GS-HS. The prediction error of this method is only slightly worse than the prediction error of GSS-EP. G-ARD does not perform well in these data sets and seems to produce high levels of over-fitting. Specifically, it produces nearly the worst prediction results among the methods that consider the grouping information. This can be related to the fact that this method lacks a hyper-parameter to specify the desired level of group sparsity or to the multiple local maxima of the type-II likelihood. Finally, the performance of BG-LASSO, is also poor in both data sets while G-LASSO provides only good prediction results in the *Kitchen* data set. A Wilcoxon test confirms that GSS-EP and GSS-MCMC perform better than the other methods being evaluated in both data sets (p-value below 5%). On the other hand, the differences in performance between these two methods are not statistically significant. The non-parametric Wilcoxon test is used in these experiments where the train and test partitions are no longer independent because it is more conservative than the Student's t-test when the assumptions made by such a test are questionable (Demšar, 2006).

	GSS-EP	GSS-MCMC	SS-EP	G-LASSO	BG-LASSO	G-HS	G-ARD
MSE	2.17±0.10	2.17±0.10	2.26±0.10	2.30±0.10	2.35±0.16	2.23±0.08	2.48±0.23
Time	5.14±1.37	4202±563	3.09±1.00	1.99±1.78	7416±513	7571±397	4.77±2.95

Table 6: MSE and average training time for each method on the *Books* data set.

	GSS-EP	GSS-MCMC	SS-EP	G-LASSO	BG-LASSO	G-HS	G-ARD
MSE	1.95±0.11	1.94±0.11	2.08±0.10	2.04±0.14	2.21±0.19	2.02±0.08	2.19±0.23
Time	4.00±0.97	2967±109	2.24±0.61	1.94±1.42	5473±278	5699±436	3.69±1.62

Table 7: MSE and average training time for each method on the *Kitchen* data set.

When we compare the average training time of the different methods we observe that GSS-EP is among the fastest methods. Nevertheless, the average training time of SS-EP, G-LASSO and G-ARD is slightly better. The reason for this is that the EP algorithm takes more iterations to achieve convergence in the two sentiment data sets considered. In any case, the results reported show that GSS-EP is still significantly faster than GSS-MCMC, BG-LASSO or G-HS, the induction methods that rely on Gibbs sampling for approximate inference. Note that in these experiments we do not report plots of the values of the model coefficients, as estimated by each different method, because the differences among these cannot be appreciated by visual inspection. Furthermore, unlike in the previous experiments about the reconstruction of sparse signals, in these data sets the optimal values of the model coefficients are unknown.

In this section we also evaluate the utility of sequential experimental design to select the training instances to induce the different models. Specifically, we repeat the previous experiments when considering an iteratively increasing number of training instances for induction and report the resulting prediction performance of G-ARD, GSS-EP and SS-EP. For each data sets, that is, *Books* and *Kitchen*, we consider the train and test partitions of the data previously generated. Initially, we train each model with the set of 100 instances contained in training set. Then, we iteratively select one by one up to 64 additional training instances from a validation set. This validation set contains 256 instances randomly extracted from the test set. The instances selected are then iteratively included in the training set, which is used again for induction. Once each model has been re-trained, we report the corresponding average prediction error on the remaining test data. The training instance that is extracted from the validation set and included in the training set is chosen using the criterion for sequential experimental design described in Section 4. Namely, we select the instance $(\mathbf{x}_{\text{new}}, y_{\text{new}})$ that maximizes $\mathbf{x}_{\text{new}}^{\text{T}} \mathbf{V} \mathbf{x}_{\text{new}}$, where $\mathbf{x}_{\text{new}}$ has been normalized to have a unit ℓ_2-norm. Note that the target y_{new} is not used in this process. The corresponding target can be obtained once $\mathbf{x}_{\text{new}}$ has been selected and included in the training set. To investigate the benefits of this strategy, we also report results when these new 64 instances are selected randomly from the validation set. Finally, in each train and test partition of the data we set the different hyper-parameters of each method to the same values as the ones used in the previous experiments.

The results of those experiments are displayed in Figure 6 for each sentiment data set. The curves in this figure show the average prediction performance of each method as a function of the number of additional training instances considered for induction when these instances are selected from the validation set randomly or using sequential experimental design. We note that in both sentiment data sets sequential experimental design tends to improve the prediction error with respect to random selection. Nevertheless, these improvements are generally below the ones reported in Figure 5. Specifically, in the case of SS-EP the observed improvements are marginal. The smaller gains observed in this figure can be explained by the fact that in these experiments we do not generate $\mathbf{x}_{\text{new}}$, but select it from a validation set. In the case of GSS-EP the observed improvements in the prediction error are more significant. Furthermore, there is statistical evidence indicating that sequential experimental design provides better prediction performance in GSS-EP than random selection when 164 training instances are used for induction. Specifically, a Wilcoxon t-test comparing the prediction error of sequential experimental design and the prediction error of random selection provides a p-value smaller than 5%. This illustrates again the benefits of sequential experimental design and the favorable properties of considering the grouping information during the induction process. Finally, we note that in the case of G-ARD sequential experimental design provides much larger improvements in the prediction error. Notwithstanding, the results of this method are far from the ones of GSS-EP.

7.3 Reconstruction of Images of Hand-written Digits

A last batch of experiments is carried out to assess the effectiveness of the generalized spike-and-slab prior to reconstruct images of hand-written digits extracted from the MNIST data set (LeCun et al., 1998). These images can also be interpreted as random signals $\mathbf{w}$ to be reconstructed from a small set of random measurements. The MNIST data set contains 60,000 digit images of 28×28 pixels (i.e., $d = 784$). This means that in this data set there are about 6,000 images of each different digit from 0 to 9. The images are in gray scale and each pixel takes a value between 0 and 255.

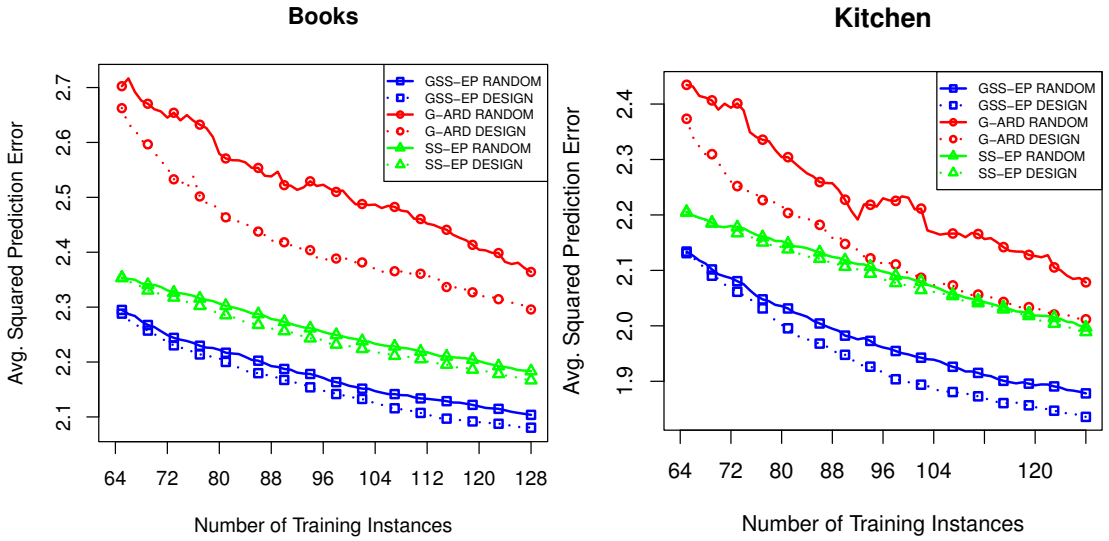

Figure 6: Mean squared prediction error for the *Books* and the *Kitchen* data sets as a function of the number of training instances for GSS-EP, G-ARD and SS-EP. We report results when the training instances are chosen randomly from a validation set and when they are selected from this data set using sequential experimental design.

Most of the pixels in each image are inactive and hence take values equal to 0. Conversely, only a few pixels are active in each image and take values near 255. Thus, these images are sparse and adequate to be reconstructed from a small set of random measurements using the method proposed.

We randomly extract 200 images of each digit from the MNIST data set. The first 100 random images are reconstructed from noisy measurements. The remaining 100 random images are used to generate the corresponding grouping information. Specifically, for each digit we generate $784/4 = 196$ groups of 4 pixels as follows.[3] First, we randomly select an initial pixel to represent each group. Then, for each different group we iteratively add, from the remaining set of pixels, the pixel that has on average the most similar activation pattern with respect to the pixels already included in the group. Similarity is measured in terms of the ℓ_2-norm. This process is repeated until all groups contain exactly 4 pixels. The second set of 100 images corresponding to each different digit are exclusively used to estimate the similarity between the activation patterns of the pixels. Once the different groups of pixels have been generated we proceed to reconstruct the set of 100 random images of each digit. The images are first normalized so that each pixel takes a value in the interval $[0,1]$ (we divide each pixel value by 255). Then, each image is stored in the vector $\mathbf{w}_0$ of $d = 784$ model coefficients. This vector contains the image to be reconstructed. Given a particular image $\mathbf{w}_0$ we then generate a reduced amount of measurements using a design matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathrm{T}}$ whose rows are sampled uniformly in the hyper-sphere of radius $\sqrt{d}$. For the reconstruction of each image, a total of $n = 288$ noisy measurements $\mathbf{y} = \{y_1, \ldots, y_n\}$ are used. This particular number of measurements is chosen because it allows to accurately reconstruct the images considered and also to discriminate among the different methods being compared (Figure 9 also reports results

3. This particular choice of the number of groups and their size is supported by the good performance results obtained by the different methods that use the grouping information.

for other numbers of measurements). Finally, each measurement is generated as $y_i = \mathbf{w}_0^{\mathrm{T}}\mathbf{x}_i + \varepsilon_i$, for $i = 1,\ldots,n$, where ε_i follows a standard Gaussian distribution, that is, the noise is introduced artificially.

Given $\mathbf{X}$ and $\mathbf{y}$ we induce $\mathbf{w}_0$ using the different methods evaluated. Let $\hat{\mathbf{w}}$ be the corresponding estimate of the image $\mathbf{w}_0$. The reconstruction error is measured as $||\hat{\mathbf{w}} - \mathbf{w}_0||_2/||\mathbf{w}_0||_2$, where $||\cdot||_2$ denotes the ℓ_2-norm. In these experiments the hyper-parameters of each method are fixed optimally. In GSS-EP and GSS-MCMC p_0, that is, the fraction of groups initially expected to be relevant for prediction, is set for each different image equal to the actual fraction of groups whose associated pixels take values different from zero. For these two methods, v_0, that is, the variance of the slab, is set equal to the square of the average deviation from zero of each component of $\mathbf{w}_0$ that is different from zero. In SS-EP the same value is used for v_0, but p_0 is set equal to the actual fraction of pixels that take values different from zero. In BG-LASSO we set γ so that the resulting prior has the same variance as the generalized spike-and-slab prior in GSS-EP and GSS-MCMC. In G-LASSO we try different values for k and report the best performing value observed for each different digit. In G-HS we fix τ so that the marginals under the group horseshoe prior have the same distance between the percentiles 1% and 99% as the marginals under the generalized spike-and-slab prior. Finally, in all methods except G-LASSO, σ_0^2, that is, the variance of the Gaussian noise, is set equal to one.

The results of these experiments are displayed in Table 8. This table shows the average reconstruction error of each method for each different digit of the MNIST data set. The average training time in seconds of each method is also displayed on the last row of the table. We have high-lighted in bold face the reconstruction error of each method when it is better than the reconstruction error of GSS-EP and there is statistical evidence indicating a performance difference. A paired t-test has been used for this purpose (p-value $< 5\%$). Similarly, when there is not enough statistical evidence to indicate a performance difference (GSS-EP performs similarly), the corresponding error has been underlined. The results where GSS-EP is found to perform better according to a t-test have been left un-modified. The table shows that the best reconstruction errors are obtained by GSS-MCMC closely followed by GSS-EP. As a matter of fact for three digits, that is, 1, 3 and 7, there is no performance difference between the two methods. GSS-EP offers a better reconstruction error than the other methods that use the grouping information. These differences in the reconstruction error are statistically significant, except when comparing results with those of G-ARD for the digits 4 and 6. In those cases there is not enough statistical evidence to indicate a performance difference. When comparing results with respect to SS-EP we observe also better reconstruction errors. This highlights the beneficial properties of considering the grouping information in the reconstruction problem. After GSS-EP and GSS-MCMC, the best performing method that uses the grouping information is G-ARD followed by G-HS. Finally, G-LASSO does not perform very well in this task and BG-LASSO obtains the worst reconstruction errors. This validates the results of Section 6.

When comparing the average training time of the different methods displayed in Table 8 we observe again that GSS-EP is among the fastest methods. In particular, GSS-EP needs only 85 seconds for training, on average. This method is again significantly faster than GSS-MCMC, BG-LASSO or G-HS, the induction methods that rely on Gibbs sampling for approximate inference. We also note that the standard deviation estimate of GSS-EP is very large. This is a consequence of the fact that the EP algorithm in some extreme situations requires a high number of iterations to achieve convergence. In any case, the training time of GSS-EP is better than the training time of SS-EP. This shows that the EP algorithm in SS-EP requires even more iterations to converge. Finally, the

	Digit	GSS-EP	GSS-MCMC	SS-EP	G-LASSO	BG-LASSO	G-HS	G-ARD
Reconstruction Error	**0**	0.26±0.14	**0.23±0.09**	0.85±0.17	0.48±0.12	0.73±0.02	0.30±0.12	0.29±0.18
	1	0.12±0.02	0.12±0.02	0.15±0.14	0.36±0.06	0.62±0.05	0.13±0.02	0.14±0.02
	2	0.24±0.14	**0.22±0.10**	0.76±0.29	0.46±0.12	0.73±0.03	0.29±0.12	0.28±0.16
	3	0.20±0.09	0.20±0.08	0.73±0.32	0.42±0.12	0.71±0.03	0.25±0.10	0.24±0.14
	4	0.18±0.07	**0.17±0.04**	0.52±0.36	0.38±0.10	0.70±0.04	0.20±0.06	0.19±0.06
	5	0.22±0.13	**0.19±0.07**	0.59±0.35	0.41±0.10	0.71±0.03	0.24±0.09	0.24±0.14
	6	0.23±0.17	**0.21±0.14**	0.69±0.32	0.42±0.14	0.71±0.04	0.26±0.14	0.24±0.17
	7	0.17±0.07	0.16±0.04	0.52±0.37	0.38±0.09	0.69±0.04	0.19±0.06	0.19±0.08
	8	0.24±0.15	**0.21±0.10**	0.79±0.25	0.45±0.13	0.72±0.03	0.28±0.13	0.29±0.20
	9	0.18±0.11	**0.17±0.08**	0.58±0.37	0.39±0.11	0.70±0.04	0.21±0.09	0.20±0.12
	Time	85±100	20730±1472	211±154	292±191	25188±1495	25536±1842	59±36

Table 8: Average reconstruction error of each method on each digit of the MNIST data set and average training time in seconds.

training time of G-ARD is also small and slightly better than the training time of GSS-EP while the training time of G-LASSO is slightly longer.

Figure 7 displays in gray scale, for a given instance of the image reconstruction problem, the actual image to be reconstructed $\mathbf{w}_0$ and the corresponding reconstructions generated by each method. The figure shows that GSS-EP and GSS-MCMC obtain the most accurate reconstructions and only include some pixels with values slightly different from zero, probably due to the measurement noise. Furthermore, the reconstructions from these two methods look nearly identical. This gives further evidence indicating that the EP posterior approximation of the posterior mean is accurate. SS-EP completely fails to reconstruct the original image (except in the images corresponding to the digit 1) and actually includes in the solution several pixels that only explain the observed data by chance. Similarly, many important pixels are excluded from the solution found by SS-EP. The reconstructions generated by G-HS and G-ARD look accurate for some images. However, these methods include many coefficients with values just slightly different from zero which were not present in the original image. The images reconstructed by G-LASSO also include many coefficients that take values slightly different from zero. Furthermore, this method tends to underestimate the values of some pixels. Again, this behavior is due to the properties of the corresponding Multi-variate Laplace prior which is not able to achieve high levels of sparsity without shrinking too much the model coefficients that are different from zero. Finally, the image reconstructed by BG-LASSO is not sparse at all which questions again the utility of the Bayesian group LASSO for group feature selection.

In these experiments we also analyze the utility of the posterior approximation computed by EP in GSS-EP to identify relevant features for prediction. In particular, for each different digit and image, we record the value of the parameters $\sigma(p_g) \in [0,1]$, with $g = 1,\ldots,G$, of the EP approximation to $\mathcal{P}(\mathbf{z}|\mathbf{y},\mathbf{X})$. This approximate distribution is described in detail in (7) and each parameter $\sigma(p_g)$ estimates the posterior probability of using the g-th group of pixels for prediction. Thus, we can simply estimate the probability of using a particular pixel for the reconstruction task by looking at the corresponding parameter $\sigma(p_g)$. To evaluate the benefits of considering the grouping information for the computation of the pixel importance, we also analyze the results obtained by SS-EP. In SS-EP there is a group for each different pixel and hence $\sigma(p_g)$, with $g = 1,\ldots,d$, directly esti-

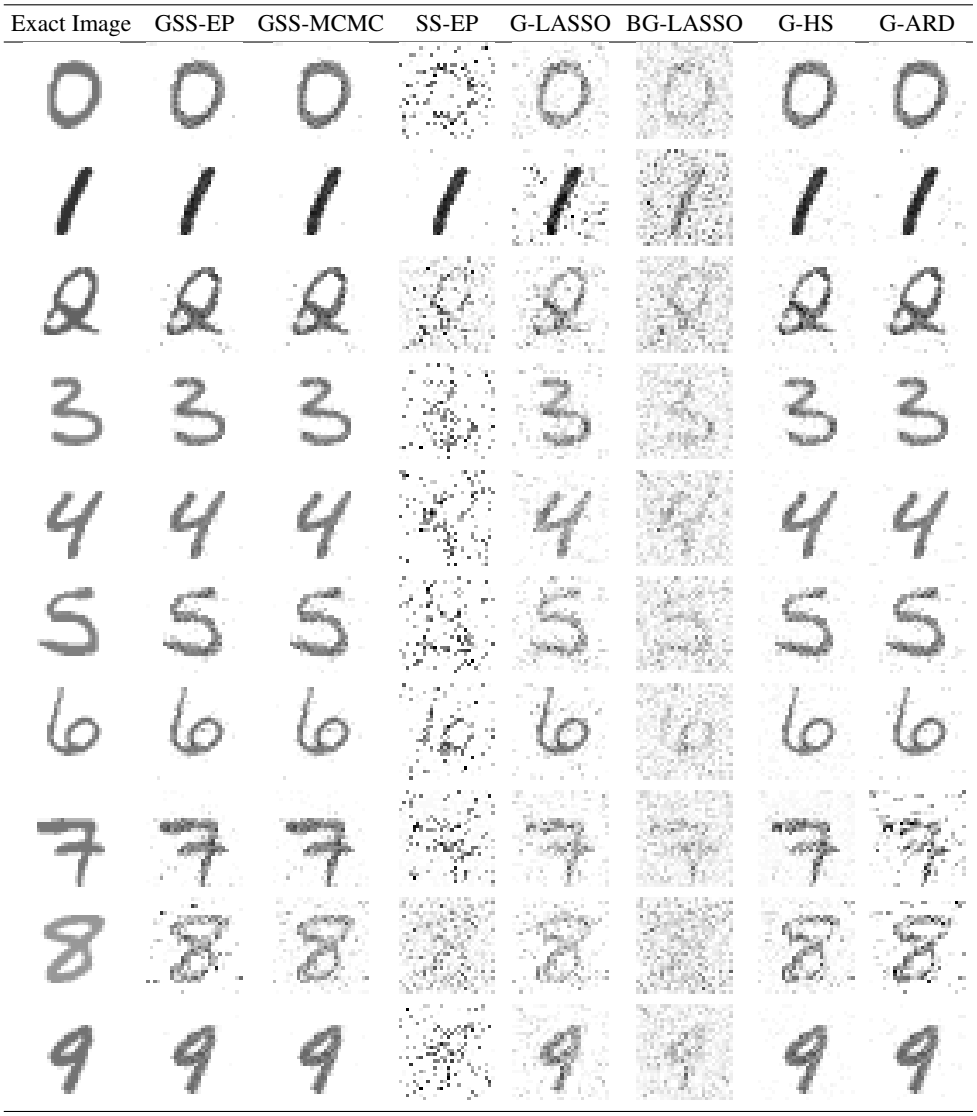

Figure 7: Representative exact images from the MNIST data set in gray scale for each different digit and the corresponding reconstructions obtained by each different method evaluated.

mates the corresponding pixel importance. Additionally, to compare results with the method based on Gibbs sampling we also report the corresponding estimates of GSS-MCMC. In GSS-MCMC the pixel importance is computed by estimating the average fraction of times the prior variance of the corresponding coefficient is different from zero within the generated samples (see Appendix A for further details). The results of these analyses are displayed graphically in Figure 8. This figure shows the importance of each pixel for the image reconstruction as estimated by GSS-EP, GSS-MCMC and SS-EP for each different digit. The results are averages over the 100 different images corresponding to each different digit. The importance of each pixels is displayed using a gray scale from 0, which corresponds to white, to 1, which corresponds to black. The results displayed show

that GSS-EP is very confident in the estimation of the pixel importance: most pixels are unlikely to be used for the reconstruction problem. Conversely, only a few pixels are very likely to be used for this task. In the case of SS-EP the results displayed show that this method is much less confident about the relative importance of each pixel for the image reconstruction. Specifically, most pixels show an intermediate level of importance, except for those of digit 1. This is probably due to the larger level of sparsity of these images, where only a few measurements are needed to identify the actual sparsity pattern. Finally, GSS-EP and GSS-MCMC offer nearly identical results. This gives further evidence supporting the accuracy of the posterior approximation computed by EP. We do not report here results for the other methods evaluated because they do not directly provide and estimate of the relative importance of each group of features for prediction.

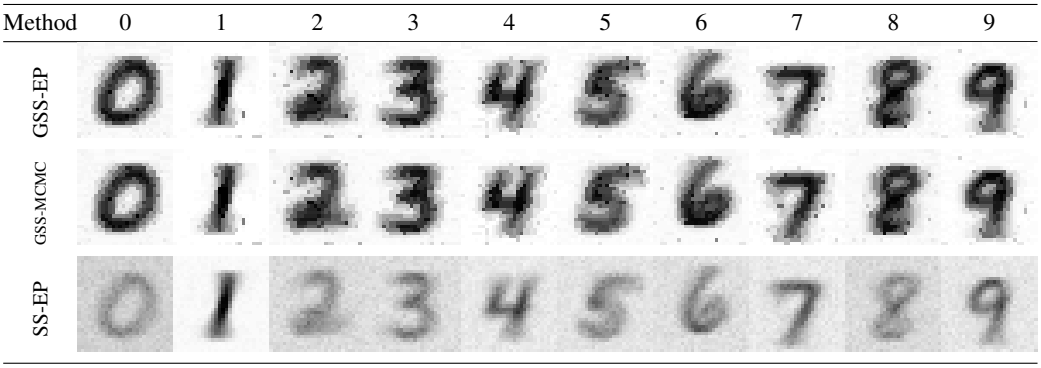

Figure 8: Average importance of each pixel for the reconstruction of the images corresponding to each different digit, as estimated by GSS-EP, GSS-MCMC and SS-EP. The feature importance of each pixel is represented in a gray scale from 0, which corresponds to the white color, to 1, which corresponds to the black color.

We also evaluate the utility of sequential experimental design to generate the different measurements used for the reconstruction of the images of hand-written digits. For this, we repeat the previous experiments for an iteratively increasing number of measurements and report the reconstruction error of each method. Again, we focus on GSS-EP, SS-EP and G-ARD. In these experiments, we start from an initial set of 128 measurements which are randomly generated using a design matrix $\mathbf{X}$ whose rows are sampled uniformly in the hyper-sphere of radius $\sqrt{d}$. We then iteratively generate up to 160 extra measurements and report the corresponding reconstruction error. The new measurements are generated using the sequential experimental design strategy described in Section 4. Finally, we also report results when these new 160 measurements are randomly generated.

The result of these experiments are displayed in Figure 9. The figure shows the reconstruction error of GSS-EP, SS-EP, and G-ARD as a function of the number of measurements performed when $\mathbf{X}$ is designed or randomly chosen for a representative subset of the ten digits contained in the MNIST data set (similar results are obtained for the digits not shown). Again, the curves displayed indicate that sequential experimental design significantly improves the reconstruction error when compared to random design. In particular, a steeper error descent with respect to the number of measurements made is produced for GSS-EP, SS-EP and G-ARD. The smallest final reconstruction error is achieved by GSS-EP followed by G-ARD, when $\mathbf{X}$ is designed. This illustrates the benefits of sequential experimental design. The reconstruction error of GSS-EP is also much better than the

reconstruction error of SS-EP when $\mathbf{X}$ is designed. This high-lights again the beneficial properties of considering the grouping information for the induction process.

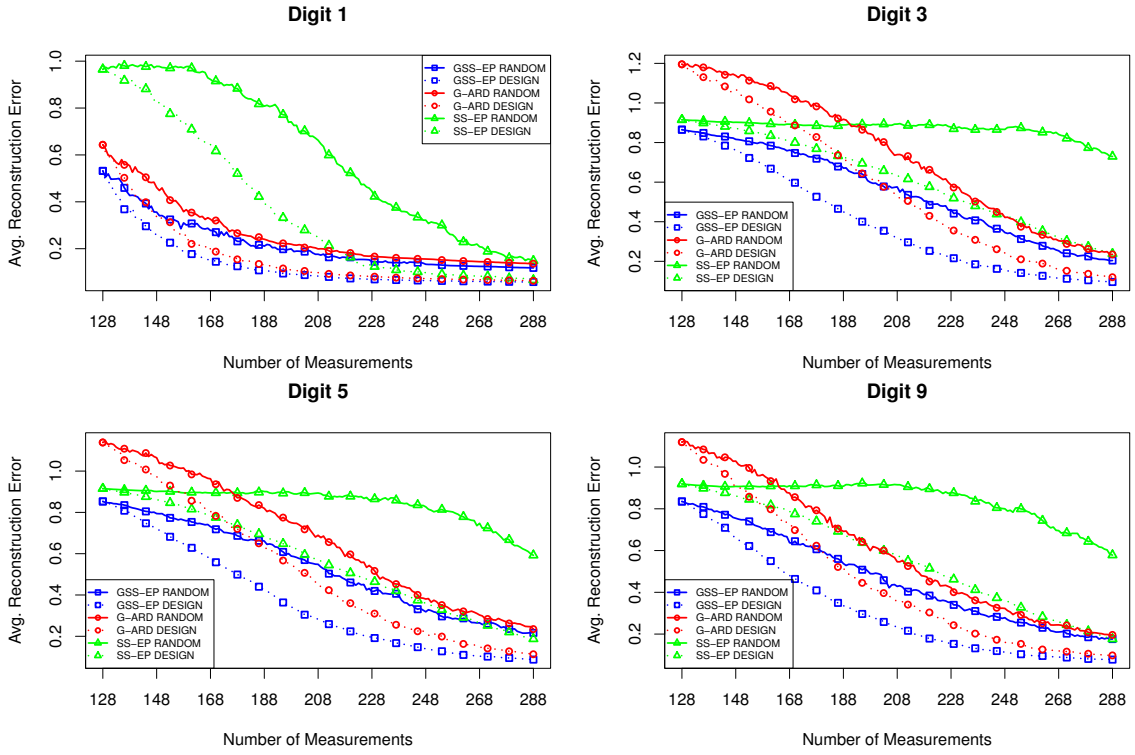

Figure 9: Reconstruction error for a representative subset of the ten digits contained in the MNIST data set as a function of the number of measurements carried out. We report results for GSS-EP, G-ARD and SS-EP when the measurements are chosen randomly and when they are chosen using sequential experimental design.

Finally, in this section we evaluate the utility of the generalized spike-and-slab prior to favor/penalize specific groups of features which are *a priori* believed to be more relevant/irrelevant for prediction. In particular, the prior distribution for $\mathbf{z}$, as defined in (4), allows to specify a different prior probability $p_{0,g}$, with $g = 1, \ldots, G$, of using the g-th group of features for prediction. In a typical application, all these parameters are set equal to a constant p_0. However, when there is prior information before hand about the relevancy or irrelevancy of specific group of features, this can be easily codified in GSS-EP by modifying these parameters. To evaluate this characteristic of the generalized spike-and-slab prior, we repeat the previous experiments where we report the reconstruction error of GSS-EP as a function of the number of random measurements performed, from 128 to 288 measurements. In these experiments we compare the results of GSS-EP when no specific prior information is used and when this information is actually used in the induction process. For this, we consider the 100 images that were used to identify each group of features. For each different digit, we evaluate the average ℓ_2-norm of the pixels contained in each group. Then, we double the $p_{0,g}$ parameter for the 25 groups with the largest estimated ℓ_2-norm. Similarly, we

reduce by half the $p_{0,g}$ parameter for the 25 groups with the smallest estimated ℓ_2-norm. Recall that initially we set all these parameters equal to the actual fraction of groups that were relevant for prediction. To further evaluate the properties of the mechanism described to introduce prior information in the induction process, we also report results when we randomly reduce by half or double the value of the parameter $p_{0,g}$ for 50 randomly chosen groups. Such a protocol evaluates the effect of wrongly choosing the prior information. We do not evaluate here the other methods described for group feature selection because they do not allow to introduce this type of prior information during the induction process.

The results of these experiments are displayed graphically in Figure 10. This figure shows the reconstruction error of GSS-EP as a function of the number of random measurements performed for a representative subset of the ten digits contained in the MNIST data set (similar results are obtained for the digits not shown). The curves report the corresponding reconstruction errors when no prior information is actually used, when the prior information is used during the induction process, and when this information is chosen randomly. The figure shows that using the prior information for induction has a beneficial effect. In particular, it leads to a significant decrease of the reconstruction error in GSS-EP. The improvements obtained are more significant when the number of measurements used for induction is small. When the prior information is chosen randomly, we observe only a slight increase of the reconstruction error of GSS-EP, which is very small compared to the gains obtained when the prior information is correctly specified. In fact, it is difficult to visually differentiate from the results obtained when no prior information is actually used. Thus, introducing prior information can be certainly beneficial for the reconstruction error, whenever such information is correct. By contrast, it barely affects the resulting model when the information is inaccurate.

8. Conclusions

In this document we have described a method for carrying out feature selection at the group level in linear regression problems. This method is able to use prior information about groups of features that are expected to be jointly relevant or irrelevant for prediction. More precisely, it is based on a linear model that considers a generalized spike-and-slab prior for group feature selection. Specifically, under this prior a set of binary latent variables is introduced, one for each different group of features, and each latent variable indicates whether or not the corresponding group is used for prediction. Exact inference under this prior is infeasible in typical regression problems. However, expectation propagation (EP) can be used as a practical alternative to carry out approximate Bayesian inference. The computational cost of EP is in $O(n^2 d)$, where n is the number of training instances and d is the number of features. This linear cost with respect to d is very efficient when $n \ll d$, which is the typical scenario we consider. Furthermore, the EP approximation provides an estimate of the posterior covariances of $\mathbf{w}$, that is, the vector of model coefficients. These covariances are shown to be very useful to carry out sequential experimental design in the linear regression model. In particular, they can be used to determine which instance to include in the training set to obtain the most information about $\mathbf{w}$, saving on costly experiments. The generalized spike-and-slab prior is also shown to be very useful to introduce prior knowledge about specific groups of features that are *a priori* expected to be more relevant or more irrelevant for prediction than the other groups. When this information is not available, the prior considered has only two hyper-parameters: p_0 and v_0. The interpretation of these parameters is very intuitive. They respectively describe the prior fraction of groups expected to be relevant for prediction and the prior variance of the model coefficients of

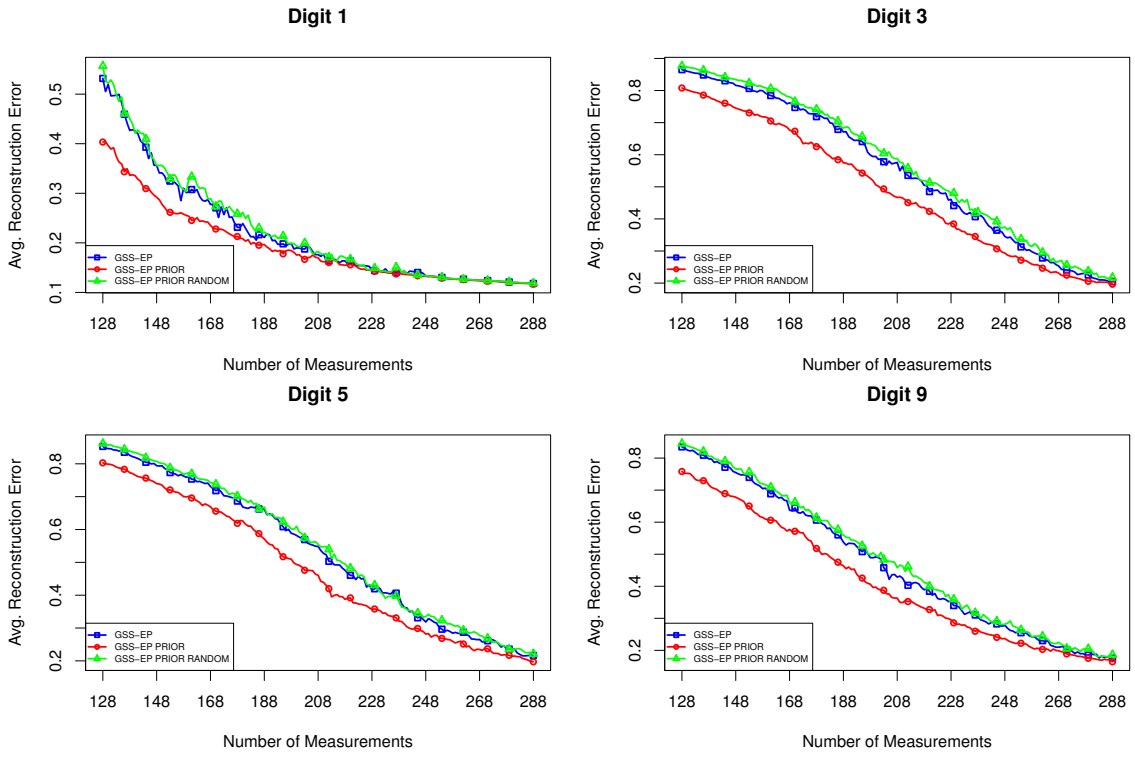

Figure 10: Reconstruction error for a representative subset of the ten digits contained in the MNIST data set as a function of the number of measurements carried out. We report results for GSS-EP when the measurements are chosen randomly. Furthermore, we consider different cases of prior information: (i) no prior information is used about the relevancy of each group of features (GSS-EP), (ii) the prior information is actually employed (GSS-EP PRIOR), and (iii) the prior information is chosen randomly (GSS-EP PRIOR RANDOM).

the relevant groups. Thus, unlike in other methods for group feature selection, in the generalized spike-and-slab prior it is very easy to specify the expected level of group sparsity and the expected deviation from zero of the relevant coefficients. If this information is available beforehand, or can be deduced from additional data, it can be readily introduced in the prior. Finally, the proposed method has the advantage of providing a posterior estimate of the importance of each group of features. This estimate can be used to identify the most relevant groups.

A detailed analysis compares the regularization properties of the generalized spike-and-slab prior considered in this document with the properties of the priors used in other methods that can also be used for group feature selection: namely, the group LASSO, the Bayesian group LASSO, the group horseshoe and the group ARD principle. This analysis shows that the generalized spike-and-slab prior is very effective for group feature selection. In particular, it is the only prior that can put a positive probability mass at the origin for the coefficients corresponding to the different groups. This probability is specified by the hyper-parameter p_0. This hyper-parameter determines the sparsity at the group level and mainly affects the regularization of irrelevant coefficients for prediction. The

smaller its value, the stronger the regularization of the model coefficients that actually take small values. By contrast, the regularization of the coefficients that are actually relevant for prediction are barely affected by p_0. These coefficients are fully regularized by the second hyper-parameter of this prior, v_0. The larger its value, the smaller the regularization of these coefficients and, as in the previous case, v_0 has little impact on the coefficients that are irrelevant for prediction. In summary, under the generalized spike-and-slab prior it is possible to provide solutions that are sparse at the group level in a selective manner. More precisely, under this prior we can model very high levels of sparsity at the group level (small values of p_0), while at the same time allowing for model coefficients that are significantly different from zero (large values of v_0). This is not possible, for example, in the case of the group LASSO or the Bayesian group LASSO. The group horseshoe also enjoys this selective shrinkage property. However, the resulting prior does not have a closed form convolution with the Gaussian distribution which makes difficult to apply the EP algorithm for fast approximate inference.

An extensive collection of experiments which considers real and synthetic data sets compares the performance of a model based on the generalized spike-and-slab prior and the EP algorithm with the other methods for group feature selection. In these experiments we also compare results when Gibbs sampling is used for approximate inference instead of EP. A model which does not use the grouping information for induction is also included in the comparison. Our results indicate that when accurate prior information about relevant or irrelevant groups of features for prediction is available, group feature selection significantly improves the results of single feature selection. In addition, from the models that use the grouping information for induction, the model based on the generalized spike-and-slab prior is shown to perform best. Furthermore, the performance of this model is very similar when the EP algorithm is used for induction or when Gibbs sampling is used instead. This confirms the accuracy of the EP approximation of the posterior distribution. Additionally, the computational cost of EP is significantly better than the computational cost of the methods based on Gibbs sampling, including the Bayesian group LASSO and the group horseshoe. The computational cost of the EP algorithm is also similar or better than the computational cost of the methods based on the group ARD principle or the group LASSO. Our results also show the utility of the EP algorithm in the proposed model for carrying out sequential experimental design. In particular, sequential experimental design in this model provides better results for a smaller number of training instances. Finally, our experiments show the benefits of introducing information about specific groups of features that are expected to be more relevant or more irrelevant *a priori* for prediction. When this prior information is introduced in the model based on the generalized spike-and-slab prior, better prediction results are obtained. By contrast, if this information is misspecified (chosen randomly), the prediction performance of the resulting model does not vary significantly.

A practical issue with the generalized spike-and-slab prior is the selection of the model hyper-parameters p_0 and v_0, which can be difficult. However, unlike the other methods for group feature selection, the interpretation of these parameters is very intuitive and they can be easily set by hand or chosen to match some specific model properties. Initially, we considered the model evidence, as approximated by the EP algorithm, for this purpose. Nevertheless, our results indicate that this is not adequate in some situations. Specifically, if the assumptions made by the model are not satisfied in practice, the model evidence can lead to wrong decisions. The more general technique of cross-validation is suggested instead. Another possibility, which is left for future exploration, is to specify hyper-priors for the model hyper-parameters and to learn them simultaneously alongside with the model coefficients $\mathbf{w}$. This has already been considered in the standard spike-and-slab

prior when Markov chain Monte Carlo is used for approximate inference (West, 2003). The model described is also restricted to work with non-overlapping groups. Thus, future research directions can also consider dealing with groups of features that overlap or that are considered to be relevant for prediction following some particular hierarchy. Another path for future investigation includes considering non-Gaussian additive noise in the estimation of the model parameters or developing some method for extracting the grouping information from the data.

Acknowledgments

Daniel Hernández-Lobato and Pierre Dupont acknowledge support from the Spanish Dirección General de Investigación, project ALLS (TIN2010-21575-C02-02).

Appendix A.

In this section we show how to implement an efficient Gibbs sampler for the regression models based on the generalized spike-and-slab prior, the Bayesian group LASSO and the group horseshoe prior. As described in Table 2, all these models are very similar and only differ in the prior distribution assumed for λ_g^2, that is, the prior variance for the coefficients $\mathbf{w}_g$ corresponding to the g-th group of features. In the case of the group horseshoe, the latent variances λ_g^2 are also multiplied by the squared value of the hyper-parameter τ. For simplicity, we include this hyper-parameter in all the derivations of this section and assume that $\tau = 1$ for the Bayesian group LASSO and the generalized spike-and-slab prior. Consider the vector of latent variables $\boldsymbol{\lambda} = (\lambda_1^2, \ldots, \lambda_G^2)^{\mathrm{T}}$, where G is the total number of groups. The method described in this section consists in first conditionally sampling each component of $\boldsymbol{\lambda}$ from the corresponding posterior distribution given $\mathbf{X}$ and $\mathbf{y}$. Then, $\mathbf{w}$ is sampled conditioned to each sample of $\boldsymbol{\lambda}$. Once this is done, the samples of $\boldsymbol{\lambda}$ are discarded and the samples of $\mathbf{w}$ are used to approximate the expectations with respect to the posterior distribution of $\mathbf{w}$. This is the approach followed by the Gibbs sampling algorithm described by George and McCulloch (1997) and by Lee et al. (2003) for the standard spike-and-slab algorithm. However, we incorporate some characteristics of the framework introduced by Tipping and Faul (2003) to speed-up the computations.

First, we describe how to conditionally sample from the posterior of λ_g^2, for $g = 1, \ldots, G$. Assume $\mathcal{P}(\lambda_g^2)$ is the corresponding prior distribution of λ_g^2. Then, the logarithm of the posterior of $\boldsymbol{\lambda} = (\lambda_1^2, \ldots, \lambda_G^2)^{\mathrm{T}}$ is:

$$\log \mathcal{P}(\boldsymbol{\lambda}|\mathbf{X}, \mathbf{y}) = \log \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{C}) + \sum_{g=1}^{G} \log \mathcal{P}(\lambda_g^2) + \text{constant}, \tag{37}$$

where $\mathbf{C} = \sigma_0^2 \mathbf{I} + \mathbf{X}\mathbf{A}\mathbf{X}^{\mathrm{T}}$ is an $n \times n$ matrix and $\mathbf{A}$ is a $d \times d$ diagonal matrix whose entries are defined as $A_{jj} = \tau^2 \lambda_g^2$ if the j-th feature belongs to the g-th group for $j = 1, \ldots, d$. Consider now that all the components of $\boldsymbol{\lambda}$ are fixed except for a particular λ_g^2 corresponding to the g-th group of features. Furthermore, denote by $\boldsymbol{\lambda}_{-g}$ the vector $(\lambda_1^2, \ldots, \lambda_G^2)^{\mathrm{T}}$ where the g-th component has been omitted. Then,

$$\log \mathcal{P}(\lambda_g^2|\mathbf{X}, \mathbf{y}, \boldsymbol{\lambda}_{-g}) = \mathcal{L}(\boldsymbol{\lambda}_{-g}) + \mathcal{L}(\lambda_g^2), \tag{38}$$

where

$$\mathcal{L}(\boldsymbol{\lambda}_{-g}) = \log \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{C}_{-g}) + \sum_{g' \neq g} \log \mathcal{P}(\lambda_{g'}^2) + \text{constant},$$

$\mathbf{C}_{-g} = \sigma_0^2 \mathbf{I} + \mathbf{X}\mathbf{A}_{-g}\mathbf{X}^{\mathrm{T}}$ is a $n \times n$ matrix and $\mathbf{A}_{-g}$ is a $d \times d$ diagonal matrix equal to $\mathbf{A}$, except for the diagonal entries corresponding to the g-th group of features. These entries are set to zero. Assume that d_g is the size of the g-th group of features and that $\mathbf{X}_g$ is a $n \times d_g$ matrix which contains in each row, for each instance, only the features of the g-th group. Consider the $d_g \times d_g$ matrix $\mathbf{M} = \mathbf{X}_g^{\mathrm{T}} \mathbf{C}_{-g}^{-1} \mathbf{X}_g$. Denote by s_j and by $\mathbf{e}_j$ the j-th eigenvalue and the j-th eigenvector of $\mathbf{M}$, respectively. Then,

$$\mathcal{L}(\lambda_g^2) = \sum_{j=1}^{d_g} \mathcal{L}_j(\lambda_g^2) + \log \mathcal{P}(\lambda_g^2), \tag{39}$$

where

$$\mathcal{L}_j(\lambda_g^2) = \frac{1}{2} \left(\frac{q_j^2 \tau^2 \lambda_g^2}{1 + \tau^2 \lambda_g^2 s_j} - \log\left(1 + \tau^2 \lambda_g^2 s_j\right) \right), \tag{40}$$

and $q_j = \mathbf{e}_j^{\mathrm{T}} \mathbf{y}^{\mathrm{T}} \mathbf{C}_{-g}^{-1} \mathbf{X}_g$. From (38) and (39) it follows that

$$\mathcal{P}(\lambda_g^2|\mathbf{X}, \mathbf{y}, \boldsymbol{\lambda}_{-g}) \propto \left[\prod_{j=1}^{d_g} \exp(\mathcal{L}_j(\lambda_g^2)) \right] \mathcal{P}(\lambda_g^2). \tag{41}$$

We generate a Gibbs sampling of $\boldsymbol{\lambda} = (\lambda_1^2, \ldots, \lambda_G^2)^{\mathrm{T}}$ by running through all the components of this vector to generate a value for λ_g^2 according to the distribution in (41). Once the process of sampling each λ_g^2 has been completed, we consider that we have generated a single sample of $\boldsymbol{\lambda}$. This process is repeated until 11,000 samples of $\boldsymbol{\lambda}$ are generated. From these, the first 1,000 are discarded and the remaining are kept. Computing s_j and q_j for $j = 1, \ldots, d_g$ can be done very efficiently since we typically assume that d_g, that is, the size of the g-th group, is relatively small. Furthermore, instead of storing the matrices $\mathbf{C}_{-g}$ and $\mathbf{C}_{-g}^{-1}$ in memory, we exclusively work with the Cholesky decompositions of these matrices which are iteratively updated for each different value of g in time $O(d_g n^2)$ (Gill et al., 1974). In particular, $\mathbf{C} = \mathbf{C}_{-g} + \tau^2 \lambda_g^2 \mathbf{X}_g \mathbf{X}_g^{\mathrm{T}}$ and $\mathbf{C}_{-g} = \mathbf{C} - \tau^2 \lambda_g^2 \mathbf{X}_g \mathbf{X}_g^{\mathrm{T}}$. Those represent d_g rank-one updates of $\mathbf{C}_{-g}$ and $\mathbf{C}$, respectively. To avoid numerical errors, the Cholesky decompositions are recomputed from scratch each time ten new Gibbs samples of $\boldsymbol{\lambda}$ are generated. Finally, since $\sum_{g=1}^{G} d_g = d$, the cost of generating one sample of $\boldsymbol{\lambda}$ is $O(n^2 d)$, where d is the total number of features.

Sampling from the distribution described in (41) for $g = 1, \ldots, G$ is straight-forward in the case of the generalized spike-and-slab prior as a consequence of the simpler form of $\mathcal{P}(\lambda_g^2)$. For the Bayesian LASSO and the group horseshoe, the method described by Damien et al. (1999) is used. For this, we sample d_g auxiliary latent variables u_j, with $j = 1, \ldots, d_g$ so that $\exp(u_j) \sim \mathcal{U}[0, \exp(\mathcal{L}_j(\lambda_g^2)]$ where $\mathcal{L}_j(\cdot)$ is defined as in (40). Then, we sample λ_g^2 from $\mathcal{P}(\lambda_g^2)$ but restricted to the set $\mathcal{A}_u = \cap_{j=1}^{d_g} \mathcal{A}_{u_j}$, where $\mathcal{A}_{u_j} = \{\lambda_g^2 : \mathcal{L}_j(\lambda_g^2) > u_j\}$. Note that each function $\mathcal{L}_j(\cdot)$ has a single global maximum (Faul and Tipping, 2001), which is found at zero when $q_j^2 < s_j$, and at

$(q_j^2 - s_j)/(\tau^2 s_j^2)$ otherwise. Let $\tilde{\lambda}_g^2(j)$ be the global maximum of $\mathcal{L}_j(\cdot)$. Each set $\mathcal{A}_{u_j}$ can be identified by finding the roots of $\mathcal{L}_j(\lambda_g^2) - u_j$ in the intervals $[0, \tilde{\lambda}_g^2(j)]$ and $[\tilde{\lambda}_g^2(j), \infty]$. If $\mathcal{A}_u$ is found to be empty (this rarely occurs in practice) we do not sample a new value for λ_g^2 and use the previous one. Finally, in the group horseshoe and the Bayesian group LASSO, $\boldsymbol{\lambda}$ is initialized to a vector whose components are all equal to 1 as described by Scott (2010). In the sampler for the model based on the generalized spike-and-slab prior, $\boldsymbol{\lambda}$ is initialized to contain only $\rho_0 G$ components different from zero and equal to v_0. The chosen components are those with the smallest least squares training error.

Given the samples of the vector $\boldsymbol{\lambda}$, we sample from the conditional posterior of $\mathbf{w}$ to approximate the posterior distribution of this vector. Conditioning on $\boldsymbol{\lambda}$, we can sample $\mathbf{w}$ from a Gaussian distribution with covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\lambda}} = \mathbf{A} + 1/\sigma_0^2 \mathbf{X}^{\mathrm{T}}\mathbf{X}$ and mean vector $1/\sigma_0^2 \boldsymbol{\Sigma}_{\boldsymbol{\lambda}} \mathbf{X}\mathbf{y}$. When $d \gg n$, this procedure has a cost in $O(n^2 d)$. See Appendix B.2 of Seeger (2008). The total cost of Gibbs sampling is in $O(kn^2 d)$, where k is the number of samples to be generated from the posterior distribution. However, often $k \gg d$ for accurate inference.

Appendix B.

In this section we show how to implement the group ARD method for group feature selection. Section 5.4 shows that this method consists in finding the maximum *a posteriori* (MAP) solution of α_g, for $g = 1, \ldots, G$, where α_g is the inverse of the prior Gaussian variance of $\mathbf{w}_g$, that is, the model coefficients of the g-th group, and G is the total number of groups. A uniform prior for each α_g is assumed in this process. For simplicity, we use here the notation of Appendix A and work with $\lambda_g^2 = \alpha_g^{-1}$. Consider the vector of latent variables $\boldsymbol{\lambda} = (\lambda_1^2, \ldots, \lambda_G^2)^{\mathrm{T}}$. From (37) we have that

$$\log \mathcal{P}(\boldsymbol{\lambda}|\mathbf{X}, \mathbf{y}) = \log \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{C}) + \text{constant}, \tag{42}$$

where $\mathbf{C} = \sigma_0^2 \mathbf{I} + \mathbf{X}\mathbf{A}\mathbf{X}^{\mathrm{T}}$ is a $n \times n$ matrix and $\mathbf{A}$ is a $d \times d$ diagonal matrix whose entries are defined as $A_{jj} = \lambda_g^2$ if the j-th feature belongs to the g-th group for $j = 1, \ldots, d$. To find the maximum of (42) we use a coordinate ascent method. Consider that all the components of $\boldsymbol{\lambda}$ are fixed except for a particular λ_g^2 corresponding to the g-th group of features. Furthermore, denote by $\boldsymbol{\lambda}_{-g}$ the vector $(\lambda_1^2, \ldots, \lambda_G^2)^{\mathrm{T}}$, where the g-th component has been omitted. From (38) and (39), it follows that

$$\log \mathcal{P}(\lambda_g^2|\mathbf{X}, \mathbf{y}, \boldsymbol{\lambda}_{-g}) = \sum_{j=1}^{d_g} \mathcal{L}_j(\lambda_g^2) + \text{constant}, \tag{43}$$

where d_g is the number of features in the g-th group and $\mathcal{L}_j(\lambda_g^2)$ is defined as in (40) for $\tau^2 = 1$. We can optimize (42) with respect to $\boldsymbol{\lambda}$ by iteratively optimizing (43) for $g = 1, \ldots, G$. However, unlike the single feature selection method (Faul and Tipping, 2001), the maximum of this function has no analytical solution (Ji et al., 2009). Thus, we have to resort to non-linear optimization to find the location of the maximum. For this, the gradient of $\mathcal{L}_j(\lambda_g^2)$ with respect to λ_g^2 can be very useful:

$$\frac{d\mathcal{L}_j(\lambda_g^2)}{d\lambda_g^2} = \frac{1}{2} \sum_{j=1}^{d_g} \left(\frac{q_j^2 - s_j - \lambda_g^2 s_j^2}{(1 + \lambda_g^2 s_j)^2} \right),$$

where q_j and s_j are defined as in (40).

In our implementation each λ_g^2 is initialized to 0. Furthermore, as in Appendix A, instead of storing the matrices $\mathbf{C}_{-g}$ and $\mathbf{C}_{-g}^{-1}$ in memory, we exclusively work with the Cholesky decompositions of these matrices which are iteratively updated for each different value of g in time $O(d_g n^2)$

(Gill et al., 1974). We take care of doing the updates of these matrices only when they are strictly needed. For example, if the optimal λ_g^2 is found to be equal to zero then $\mathbf{C} = \mathbf{C}_{-g}$. The total cost of the algorithm described is in $O(n^2 d)$ under the assumption that the number of relevant groups is in $O(d)$, where d is the total number of features.

References

H. Attias. A variational Bayesian framework for graphical models. In *In Advances in Neural Information Processing Systems*, volume 12, pages 209–215. MIT Press, 2000.

F. R. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008. ISSN 1532-4435.

J. A. Bazerque, G. Mateos, and G. B. Giannakis. Group-lasso on splines for spectrum cartography. *Transactions on Signal Proccesing*, 59(10):4648–4663, October 2011. ISSN 1053-587X.

C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, August 2006. ISBN 0387310738.

J. Blitzer, M. Dredze, and F. Pereira. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–407, Prague, Czech Republic, 2007.

E. J. Candes and M. B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21 –30, march 2008. ISSN 1053-5888.

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.

K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10: 273–304, 1995.

P. Damien, J. Wakefield, and S. Walker. Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society Series B*, 61(2): 331–344, 1999.

J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006. ISSN 1533-7928.

D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

A. C. Faul and M. E. Tipping. Analysis of sparse Bayesian learning. In Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 383–389. MIT Press, 2001.

V. Fedorov. *Theory of Optimal Experiments*. Academic Press, 1972.

E. I. George and R. E. McCulloch. Approaches for Bayesian variable selection. *Statistica Sinica*, 7 (2):339–373, 1997.

M. Van Gerven, B. Cseke, R. Oostenveld, and T. Heskes. Bayesian source localization with the multivariate Laplace prior. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1901–1909, 2009.

J. Geweke. Variable selection and model comparison in regression. *Bayesian Statistics*, 5:609–620, 1996.

P. E. Gill, G. H. Golub, W. Murray, and M. A. Saunders. Methods for modifying matrix factorizations. *Mathematics of Computation*, 28(126):505–535, 1974.

D. Hernández-Lobato. *Prediction Based on Averages over Automatically Induced Learners: Ensemble Methods and Bayesian Techniques*. PhD thesis, Computer Science Department, Universidad Autónoma de Madrid, 2009. Online available at: http://arantxa.ii.uam.es/d̃hernan/docs/Thesis_color_links.pdf.

D. Hernández-Lobato, J. M. Hernández-Lobato, T. Helleputte, and P. Dupont. Expectation propagation for Bayesian multi-task feature selection. In José L. Balcázar, Francesco Bonchi, Aristides Gionis, and Michèle Sebag, editors, *Proceedings of the European Conference on Machine Learning*, volume 6321, pages 522–537. Springer, 2010.

J. M. Hernández-Lobato. *Balancing Flexibility and Robustness in Machine Learning: Semiparametric Methods and Sparse Linear Models*. PhD thesis, Computer Science Department, Universidad Autónoma de Madrid, 2010.

J. M. Hernández-Lobato and T. Dijkstra. Hub gene selection methods for the reconstruction of transcription networks. In José L. Balcázar, Francesco Bonchi, Aristides Gionis, and Michèle Sebag, editors, *Machine Learning and Knowledge Discovery in Databases, European Conference, Proceedings, Part I*, volume 6321, pages 506–521, 2010.

J. M. Hernández-Lobato, T. Dijkstra, and T. Heskes. Regulator discovery from gene expression time series of malaria parasites: A hierarchical approach. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 649–656. The MIT Press, 2008.

J. M. Hernández-Lobato, D. Hernández-Lobato, and A. Suárez. Network-based sparse Bayesian classification. *Pattern Recognition*, 44:886–900, 2011.

J. Huang and T. Zhang. The benefit of group sparsity. *The Annals of Statistics*, 38:1978–2004, 2010.

T. S. Jaakkola. Tutorial on variational approximation methods. In M. Opper and D. Saad, editors, *Advances in Mean Field Methods: Theory and Practice*, pages 129–159. MIT Press, 2001.

S. Ji and L. Carin. Bayesian compressive sensing and projection optimization. In *Proceedings of the 24th International Conference on Machine Learning*, pages 377–384. ACM, 2007. ISBN 978-1-59593-793-3.

S. Ji, Y. Xue, and L. Carin. Bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 56(6):2346 –2356, june 2008. ISSN 1053-587X.

S. Ji, D. Dunson, and L. Carin. Multitask compressive sensing. *IEEE Transactions on Signal Processing*, 57(1):92 –106, January 2009. ISSN 1053-587X.

I. M. Johnstone and D. M. Titterington. Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367 (1906):4237, 2009.

H. J. Kappen and Vicenç Gómez. The variational garrote. *Machine Learning*, pages 1–17, 2013. Submitted.

S. Kim and E. P. Xing. Feature selection via block-regularized regression. In David A. McAllester and Petri Myllymäki, editors, *Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence*, pages 325–332. AUAI Press, 2008.

Y. Kim, J. Kim, and Y. Kim. Blockwise sparse regression. *Statistica Sinica*, 16(2):375, 2006. ISSN 1017-0405.

N. Lawrence, M. W. Seeger, and R. Herbrich. Fast sparse Gaussian process methods: The informative vector machine. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 609–616. MIT Press, Cambridge, MA, 2003.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

K. E. Lee, N. Sha, E. R. Dougherty, M. Vannucci, and B. K. Mallick. Gene selection: A Bayesian variable selection approach. *Bioinformatics*, 19(1):90–97, 2003.

Y. Li, C. Campbell, and M. Tipping. Bayesian automatic relevance determination algorithms for classifying gene expression data. *Bioinformatics*, 18(10):1332–1339, 2002.

D. J. C. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4:590–604, 1991.

D. J. C. MacKay. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, Cambridge, UK., 2003. ISBN 0521642981.

D. M. Malioutov, M. Çetin, and A. S. Willsky. A sparse signal reconstruction perspective for source localization with sensor arrays. *IEEE Transactions on Signal Processing*, 53(8-2):3010–3022, 2005.

L. Meier, S. Van De Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008. ISSN 1467-9868.

T. Minka. *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, Massachusetts Institute of Technology, 2001.

T. Minka and J. Lafferty. Expectation-propagation for the generative aspect model. In Adnan Darwiche and Nir Friedman, editors, *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pages 352–359. Morgan Kaufmann, 2002.

T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988.

R. M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto, 1993.

H. Nickisch and C. E. Rasmussen. Approximations for binary Gaussian process classification. *Journal of Machine Learning Research*, 9:2035–2078, 10 2008.

B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 79–86, 2002.

A. T. Puig, A. Wiesel, G. F., and A. O. Hero. Multidimensional shrinkage-thresholding operator and group lasso penalties. *IEEE Signal Processing Letters*, 18(6):363–366, 2011.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

S. Raman, T. J. Fuchs, P. J. Wild, E. Dahl, and V. Roth. The Bayesian group-lasso for analyzing contingency tables. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 881–888, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1.

V. Roth and B. Fischer. The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In A. McCallum and S. Roweis, editors, *Proceedings of the 25th International Conference on Machine Learning*, pages 848–855, 2008.

T. Sandler, J. Blitzer, P. P. Talukdar, and L. H. Ungar. Regularized learning with networks of features. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1401–1408. 2009.

F. Scheipl, L. Fahrmeir, and T. Kneib. Spike-and-slab priors for function selection in structured additive regression models. *Journal of the American Statistical Association*, 107:1518–1532, 2012.

J. G. Scott. Parameter expansion in local-shrinkage models. *ArXiv E-prints*, 2010. arXiv:1010.5265v1.

M. W. Seeger. Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research*, 9:759–813, 2008.

M. W. Seeger, H. Nickisch, R. Pohmann, and B. Schölkopf. Optimization of k-space trajectories for compressed sensing by Bayesian experimental design. *Magnetic Resonance in Medicine*, 63(1): 116–126, 2009.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

M. E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

M. E. Tipping and A. Faul. Fast marginal likelihood maximisation for sparse Bayesian models. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, pages 3–6, 2003.

J. E. Vogt and V. Roth. The group-lasso: $\ell_{1,\infty}$ regularization versus $\ell_{1,2}$ regularization. In Goesele et al., editor, *32nd Anual Symposium of the German Association for Pattern Recognition*, volume 6376, pages 252–261. Springer, 2010.

M. Wakin, M. Duarte, S. Sarvotham, D. Baron, and R. Baraniuk. Recovery of jointly sparse signals from few random projections. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 1433–1440. MIT Press, Cambridge, MA, 2006.

M. West. Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian Statistics 7*, pages 723–732. Oxford University Press, 2003.

T. J. Yen and Y. M. Yen. Grouped variable selection via nested spike and slab priors. *ArXiv E-prints*, 2011. arXiv:1106.5837v1.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006. ISSN 1467-9868.

H. Zhu and R. Rohwer. Bayesian invariant measurements of generalization. *Neural Processing Letters*, 2:28–31, 1995. ISSN 1370-4621.