Bayesian Models for Structured Sparse Estimation via Set Cover Prior

Xianghang Liu[1,2], Xinhua Zhang[1,3], Tibério Caetano[1,2,3]

[1] Machine Learning Research Group, National ICT Australia, Sydney and Canberra, Australia
[2] School of Computer Science and Engineering, The University of New South Wales, Australia
[3] Research School of Computer Science, The Australian National University, Australia
{xianghang.liu, xinhua.zhang, tiberio.caetano}@nicta.com.au

Abstract. A number of priors have been recently developed for Bayesian estimation of sparse models. In many applications the variables are simultaneously relevant or irrelevant in groups, and appropriately modeling this correlation is important for improved sample efficiency. Although group sparse priors are also available, most of them are either limited to disjoint groups, or do not infer sparsity at group level, or fail to induce appropriate patterns of support in the posterior. In this paper we tackle this problem by proposing a new framework of prior for overlapped group sparsity. It follows a hierarchical generation from group to variable, allowing group-driven shrinkage and relevance inference. It is also connected with set cover complexity in its maximum a posterior. Analysis on shrinkage profile and conditional dependency unravels favorable statistical behavior compared with existing priors. Experimental results also demonstrate its superior performance in sparse recovery and compressive sensing.

1 Introduction

Sparsity is an important concept in high-dimensional statistics [1] and signal processing [2] that has led to important application successes. It reduces model complexity and improves interpretability of the result, which is critical when the number of explanatory variables p in the problem is much higher than the number of training instances n.

From a Bayesian perspective, the sparsity of a variable β_i is generally achieved by shrinkage priors, which often take the form of scale mixture of Gaussians: $\beta_i|z_i \sim \mathcal{N}(0, z_i)$. z_i indicates the relevance of β_i, and a broad range of priors on z_i has been proposed. For example, the spike and slab prior [3, 4] uses a Bernoulli variable for z_i, which allows β_i to be exactly zero with a positive probability. Absolutely continuous alternatives also abound [5], *e.g.*, the horseshoe prior [6, 7] which uses half-Cauchy on z_i and offers robust shrinkage in the posterior. Interestingly, the maximum a posterior (MAP) inference often corresponds to deterministic models based on sparsity-inducing regularizers, *e.g.* Lasso [8] when z_i has a gamma distribution [9, 10]. In general, the log-posterior can be non-concave [11, 12].

However, many applications often exhibit additional structures (correlations) in variables rather than being independent. Groups may exist such that variables of each group are known a priori to be jointly relevant or irrelevant for data generation. Encoding this knowledge in the prior proves crucial for improved accuracy of estimation. The

simplest case is when the groups are disjoint, and they form a partition of the variable set. This allows the relevance indicator z_i of all variables in each group to be tied, forming a group indicator which is endowed with a zero-centered prior as above [13, 14]. In particular, a gamma prior now yields the Bayesian group Lasso [15], and its MAP is the group Lasso [16] which allows group information to *provably* improve sample efficiency [17]. More refined modeling on the sparsity within each group has also been explored [18, 19]. We overview the related background in Section 2.

However, groups do overlap in many practical applications, *e.g.* gene regulatory network in gene expression data [20], and spatial consistency in images [21]. Techniques that deal with this scenario start to diverge. A commonly used class of method employs a Markov random field (MRF) to enforce smoothness over the relevance indicator of all variables within each group [22–24]. However, this approach does not infer relevance at the group level, and does not induce group-driven shrinkage.

Another popular method is to directly use the Bayesian group Lasso, despite the loss of hierarchical generative interpretation due to the overlap. Its MAP inference has also led to a rich variety of regularizers that promote structured sparsity [21, 25], although statistical justification for the benefit of using groups is no longer rich and solid. Moreover, Bayesian group Lasso tends to shrink a whole group based on a complexity score computed from its constituent variables. So the support of the posterior $\boldsymbol{\beta}$ tends to be the complement of the union of groups, rather than the union of groups as preferred by many applications.

To address these issues, we propose in Section 3 a hierarchical model by placing relevance priors on groups only, while the variable relevance is derived (probabilistically) from the set of groups that involve it. This allows direct inference of group relevance, and is amenable to the further incorporation of hierarchies among groups. All previously studied sparsity-inducing priors on relevance variables can also be adopted naturally, leading to a rich family of structured sparse prior. The MAP of our model turns out exactly the set cover complexity, which provably reduces sample complexity for *overlapped* groups [26].

Although in appearance our model simply reverses the implication of relevance in Bayesian group Lasso, it amounts to considerably more desirable shrinkage profile [7]. In Section 4, detailed analysis based on horseshoe prior reveals that set cover priors retain the horseshoe property in its posterior, shrinking reasonably for small response and diminishing when response grows. Surprisingly, these properties are not preserved by the other structured alternatives. Also observed in set cover prior is the favorable conditional dependency between relevance variables, which allows them to "explain-away" each other through the overlap of two groups they each belong to. Experimental results in Section 5 confirm that compared with state-of-the-art structured priors, the proposed set cover prior outperforms in sparse recovery and compressive sensing on both synthetic data and real image processing datasets.

Note different from [27] and [28], we do not introduce regression variables that account for interactions between features, *i.e.* β_{ij} for $x_i x_j$.

2 Preliminaries on Sparse Priors

In a typical setting of machine learning, we are given n training examples $\{\mathbf{x}_i, y_i\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^p$ represents a vector of p features/variables, and y_i is the response that takes value in $\mathbb{R}$ for regression, and in $\{-1, 1\}$ for classification. Our goal is to learn a linear model $\boldsymbol{\beta} \in \mathbb{R}^p$, or a distribution of $\boldsymbol{\beta}$, such that $\mathbf{x}_i'\boldsymbol{\beta}$ agrees with y_i. This problem is usually ill-posed, especially when $p \gg n$ as considered in this work. Therefore prior assumptions are required and here we consider a popular prior that presumes $\boldsymbol{\beta}$ is sparse. In Bayesian methods, the compatibility between y_i and $\mathbf{x}_i'\boldsymbol{\beta}$ is enforced by a likelihood function, which is typically normal for regression (*i.e.*, $y|\mathbf{x}, \boldsymbol{\beta} \sim \mathcal{N}(\mathbf{x}'\boldsymbol{\beta}, \sigma^2)$), and Bernoulli for classification. σ is a pre-specified constant.

The simplest form of sparsity is enforced on each element of $\boldsymbol{\beta}$ independently through priors on β_i. Most existing models use a scalar mixture of normals that correspond to the graphical model $z_i \to \beta_i$ [27, 29, 30]:

$$\pi(\beta_i) = \int \mathcal{N}(\beta_i; 0, \sigma_0^2 z_i) f(z_i) \mathrm{d}z_i. \tag{1}$$

Here σ_0^2 can be a constant, or endowed with a prior. Key to the model is the latent conditional variance z_i, which is often interpreted as *relevance* of the variable β_i. Larger z_i allows β_i to take larger absolute value, and by varying the mixing distribution f of z_i we obtain a range of priors on $\boldsymbol{\beta}$, differing in shrinkage profile and tail behavior. For example, the spike and slab prior [3, 4] adopts

$$f_{\mathrm{SS}}(z_i) = p_0 \delta(z_i - 1) + (1 - p_0)\delta(z_i), \tag{2}$$

where δ is the Dirac impulse function and p_0 is the prior probability that β_i is included. Absolutely continuous distributions of z_i are also commonly used. An inverse gamma distribution on z_i leads to the Student-t prior, and automatic relevance determination [ARD, 9] employs $f(z_i) \propto z_i^{-1}$. Indeed, a number of sparsity-inducing priors can be unified using the generalized beta mixture [5, 31]:

$$z_i|\lambda_i \sim \mathrm{Ga}(a, \lambda_i), \quad \text{and} \quad \lambda_i \sim \mathrm{Ga}(b, d). \tag{3}$$

Here Ga stands for the gamma distribution with shape and rate (*inverse* scale) parameters. In fact, z_i follows the generalized beta distribution of the second kind:

$$\mathrm{GB2}(z_i; 1, d, a, b) = z_i^{a-1}(1 + z_i/d)^{-a-b} d^{-a} / B(a, b), \tag{4}$$

where $B(a, b)$ is the beta function. When $a = b = \frac{1}{2}$, it yields the horseshoe prior on $\boldsymbol{\beta}$ [6]. The normal-exponential-gamma prior and normal-gamma prior [32] can be recovered by setting $a = 1$ and $b = d \to \infty$ respectively. In the intersection of these two settings is the Bayesian Lasso: $\pi(\boldsymbol{\beta}) \sim \exp(-\|\boldsymbol{\beta}\|_1)$ [10], where $\|\boldsymbol{\beta}\|_p := (\sum_i |\beta_i|^p)^{\frac{1}{p}}$ for $p \geq 1$.

To lighten notation, in the case of spike and slab we will also use z_i to represent Bernoulli variables valued in $\{0, 1\}$. So integrating over $z_i \geq 0$ with respect to the density in (2) can be interpreted as weighted sum over $z_i \in \{0, 1\}$.

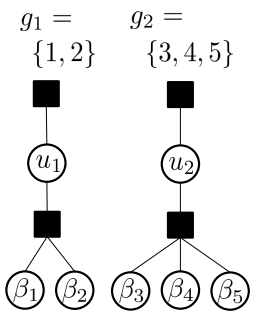


Fig. 1. Group spike and slab

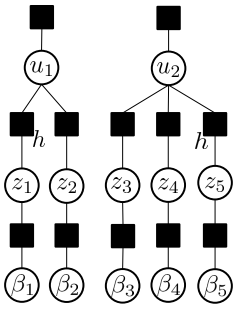


Fig. 2. Nested spike and slab

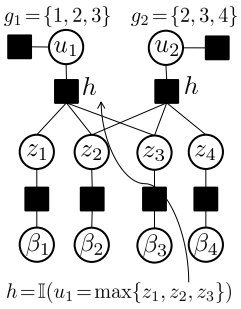


Fig. 3. Group counting prior for spike and slab

2.1 Disjoint Groups

In many applications, prior knowledge is available that the variables can be partitioned into *disjoint* groups $g_i \subseteq [p] := \{1, \ldots, p\}$, and all variables in a group tend to be positively correlated, *i.e.* relevant or irrelevant simultaneously. Denote $\mathcal{G} = \{g_1, g_2, \ldots\}$. [13] generalized the spike and slab prior to this scenario by introducing a scalar parameter of relevance for each group: $u_g \sim f_{\text{SS}}$, and extending (1) into a scalar mixture of *multivariate* normal

$$\pi(\boldsymbol{\beta}_g) = \int \mathcal{N}(\boldsymbol{\beta}_g; \mathbf{0}, \boldsymbol{\Lambda}_g u_g) f(u_g) \mathrm{d}u_g \;\; \forall\, g \in \mathcal{G}. \tag{5}$$

Here $\boldsymbol{\beta}_g$ encompasses all variables in the group g, and $\boldsymbol{\Lambda}_g$ is a diagonal matrix of variance. See Figure 1 for the factor graph representation that will facilitate a unified treatment of other models below. As a result, *correlation* is introduced among all variables in each group. Using exactly the same density f as above (but on u_g here), one may recover the group horseshoe, group ARD [14], and Bayesian group Lasso [15]:

$$\pi(\boldsymbol{\beta}) \propto \exp(-\|\boldsymbol{\beta}\|_{\mathcal{G}}), \text{ where } \|\boldsymbol{\beta}\|_{\mathcal{G}} = \sum\nolimits_g \|\boldsymbol{\beta}_g\|_p \,. \tag{6}$$

Common choices of p are 2 and ∞. To further model the sparsity of different variables within a group, [18] proposed a nested spike and slab model as shown in Figure 2. The key idea is to employ *both* Bernoulli variables z_i and u_g that encode the relevance of variables and groups respectively, and to define the spike and slab distribution of β_i conditional on $u_g = 1$. In particular, z_i must be 0 if $u_g = 0$, *i.e.* group g is excluded. This relation is encoded by a factor between z_i and u_g:

$$h(z_i, u_g) = \begin{cases} p_0^{z_i}(1-p_0)^{1-z_i} & \text{if } u_g = 1 \\ \mathbb{I}(z_i = 0) & \text{if } u_g = 0 \end{cases}, \; \forall i \in g. \tag{7}$$

Here $\mathbb{I}(\cdot) = 1$ if $\cdot$ is true, and 0 otherwise.

3 Structured Prior with Overlapped Groups

In many applications, groups may overlap and fully Bayesian treatments for this setting have become diverse.

Group counting prior (GCP). A straightforward approach is to ignore the fact of overlapping, and simply use the group Lasso prior in (6). This idea is also used in deterministic overlapped group Lasso [16]. When $p = \infty$, the norm in (6) is the Lovász extension of the group counting penalty [33] which, in the case of spike and slab prior on β_i, can be written in terms of the binary relevance indicator $\mathbf{z} := \{z_i\} \in \{0, 1\}^p$

$$\Omega(\mathbf{z}) = \prod_{g \in \mathcal{G}} p_0^{u_g} (1 - p_0)^{1 - u_g}, \quad \text{where} \quad u_g = \max_{i: i \in g} z_i. \tag{8}$$

So a group is deemed as relevant ($u_g = 1$) if, and only if, any variable in the group is relevant ($z_i = 1$). The factor graph is given in Figure 3, with a Bernoulli potential on u_g. However, since this prior promotes u_g to be 0 (*i.e.* zero out all variables in the group g), the support of $\boldsymbol{\beta}$ in the posterior tends to be the complement of a union of groups. Although this may be appropriate for some applications, the support is often more likely to be the union of groups.

MRF prior. Instead of excluding groups based on its norm, the MRF prior still places sparsity-inducing priors on each variable β_i, but further enforces *consistency* of relevance within each group via z_i. For example, assuming the variables are connected via an undirected graph where each edge $(i, j) \in E$ constitutes a group, [22, 34] extended the spike and slab prior by incorporating a pairwise MRF over the relevance indicators z_i: $\exp(-\sum_{(i,j) \in E} R_{ij} \mathbb{I}(z_i \neq z_j))$.

As a key drawback of the above two priors, they do not admit a generative hierarchy and perform no inference at the group level. To address these issues, we next construct a hierarchical generative model which explicitly characterizes the relevance of both groups and variables, as well as their conditional correlations.

3.1 Set cover prior (SCP)

To better clarify the idea, we first focus on spike and slab prior where sparsity can be easily modeled by Bernoulli variables z_i and u_g. Recall the nested model in Figure 2, where each group has a Bernoulli prior, and each variable z_i depends on the unique group that it belongs to. Now since multiple groups may be associated with each node, it will be natural to change the dependency into some arithmetics of these group indicators. In Figure 4, we show an example with[1]

$$h(z_i, \{u_g : i \in g\}) = \mathbb{I}(z_i \leq \max\{u_g : i \in g\}). \tag{9}$$

This means a variable can be relevant only if any group including it is also relevant. Although this appears simply reversing the implication relations between group and variable in the group counting prior, it does lead to a hierarchical model and enjoys much more desirable statistical properties as will be shown in Section 4.

[1] This defines a potential in an MRF; there is no explicit prior on z_i.

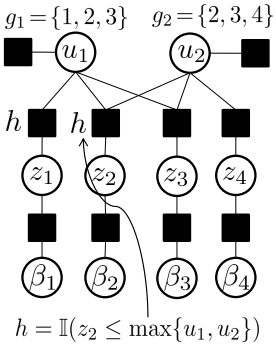


Fig. 4. Set cover prior for spike and slab

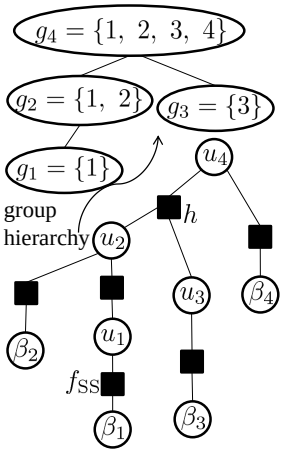


Fig. 5. Set cover prior for spike and slab with tree hierarchy. u_j corresponds to g_j. $h = \mathbb{I}(u_4 \geq \max\{u_2, u_3\})$.

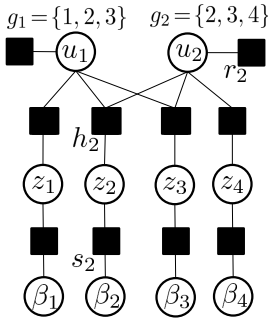


Fig. 6. Set cover prior using horseshoe. $r_2 = \text{Ga}(u_2; \frac{1}{2}, \frac{1}{2})$. $h_2 = \text{Ga}(z_2; \frac{1}{2}, \max\{u_1, u_2\})$.

By endowing a Bernoulli prior on all u_g with $\Pr(u_g = 1) = p_0 < 0.5$ (*i.e.* favoring sparsity), we complete a generative prior of $\boldsymbol{\beta}$ in a spike and slab fashion. Given an assignment of $\mathbf{z}$, it is interesting to study the mode of $\{u_g\}$, which is the solution to

$$\min_{\{u_g\}} \sum_g u_g, \quad \text{s.t. } u_g \in \{0, 1\}, \ \sum_{g: i \in g} u_g \geq z_i, \ \forall\, i. \tag{10}$$

This turns out to have exactly the same complexity as set cover [26]. It seeks the smallest number of groups such that their union covers the set of variables. Hence we will call this prior as "set cover prior". This optimization problem is NP-hard in general, and some benefit in sample complexity is established by [26].

A number of extensions follow directly. Additional priors (*e.g.* MRF) can be placed on variables z_i. The $\max$ in (9) can be replaced by $\min$, meaning that a variable can be selected only if *all* groups involving it are selected. Other restrictions such as limiting the number of selected variables in each (selected) group can also be easily incorporated [35]. Moreover, groups can assume a hierarchical structure such as tree, *i.e.* $g \cap g' \in \{g, g', \emptyset\}$ for all g and g'. Here the assumption is that if a node g' is included, then all its ancestors $g \supset g'$ must be included as well [21, 36]. This can be effectively enforced by adding a factor h that involves each group g and its children $\text{ch}(g)$ (see Figure 5):

$$h(g, \text{ch}(g)) = \mathbb{I}(u_g \geq \max\nolimits_{g' \in \text{ch}(g)} u_{g'}). \tag{11}$$

When the groups are disjoint, both set cover and group counting priors are equivalent to group spike and slab.

3.2 Extension to Generalized Beta Mixture

The whole framework is readily extensible to the continuous sparse priors such as horseshoe and ARD. Using the interpretation of z_i and u_g as relevance measures, we could

simply replace the function $\mathbb{I}$ that tests equality by the Dirac impulse function δ, and apply various types of continuous valued priors on z_i and u_g. This is indeed feasible for GCP, *e.g.* encode the continuous variant of (8) using the generalized beta mixture in (3)

$$h(u_g, \{z_i : i \in g\}) = \delta(u_g - \max\{z_i : i \in g\}), \quad h(u_g) = \mathrm{GB2}(u_g; 1, d, a, b). \tag{12}$$

Here more flexibility is available when z_i is continuous valued, because the $\max$ can be replaced by multiplication or summation, which promotes or suppresses sparsity respectively [27, Theorem 1, 2].

However problems arise in SCP if we directly use

$$z_i = \max_{g:i\in g} u_g \quad \text{or} \quad \min_{g:i\in g} u_g, \quad \text{where } u_g \sim \mathrm{GB2}(u_g; 1, d, a, b), \tag{13}$$

because it leads to singularities in the prior distribution on $\mathbf{z}$. To smooth the prior, we resort to arithmetic combinations of the *intermediate* variables in the generative process of the prior on u_g. Note that in (3), d is a scale parameter, while a and b control the behavior of the distribution of z_i close to zero and on the tail, respectively. A smaller value of λ_i places more probability around 0 in z_i, encouraging a sparser β_i. So a natural way to combine the group prior is:

$$z_i|\{u_g\} \sim \mathrm{Ga}(a, \max_{g:i\in g} u_g), \quad \text{and} \quad u_g \sim \mathrm{Ga}(b, d), \tag{14}$$

where $\max$ allows z_i to pick up the most sparse tendency encoded in all u_g of the associated groups[2]. Changing it to $\min$ leads to adopting the least sparse one. The resulting graphical model is given in Figure 6. Here u_g has a gamma distribution, playing the same role of relevance measure as in the normal-gamma prior on β_i [32]. The SCP constructed in (14) is no longer equivalent to the group priors in Section 2.1, even when the groups are disjoint.

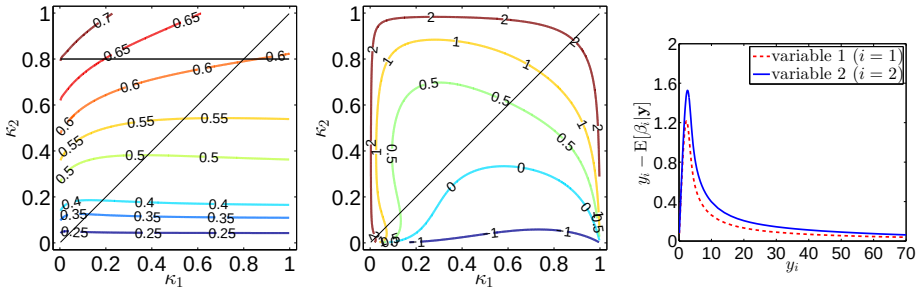
In fact, the arithmetics that combine multiple groups can be carried out at an even higher level of the generative hierarchy. For example, in the horseshoe prior where $a = b = 1/2$, one may introduce an additional layer of mixing over the scale parameter d, making it an arithmetic combination of u_g of the associated groups. We leave this possibility for future exploration.

Notice [38] used a partial least squares approach based on an MRF of binary selectors of groups and variables. However their method is confined to spike and slab, because these two groups of indicators are *not* coupled by the potential function, but by imposing external restrictions on the admissible joint assignment that is valued in $\{0, 1\}$. It also brings much challenge in MCMC inference.

4 Analysis of Structured Sparse Prior

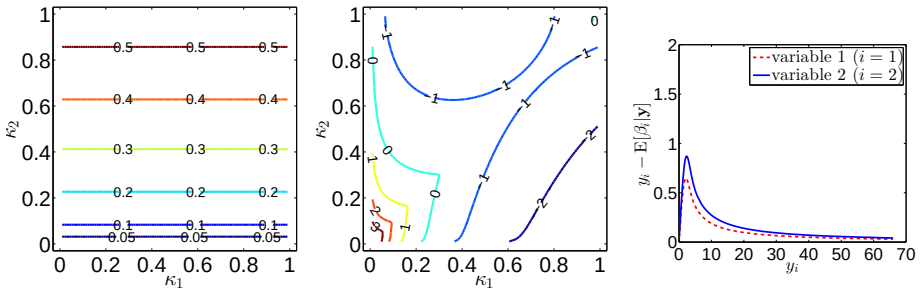
Although the above three types of priors for structured sparsity appear plausible, their statistical properties differ significantly as we study in this section. Here in addition to the robust shrinkage profile studied by [6], we also compare the conditional correlation among variables when the groups overlap.

[2] See more detailed discussions in Appendix A of the full paper [37] on how a greater value of the second argument (rate, *i.e.* inverse scale) of a Gamma distribution induces higher sparsity in β_i.



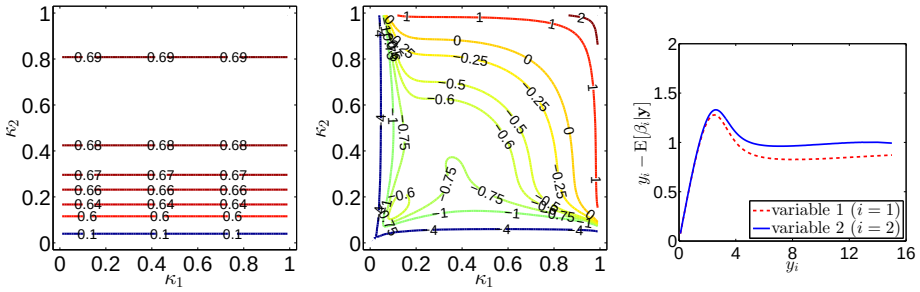
(a) Contour of $\mathbb{E}[\kappa_3|\kappa_1, \kappa_2]$ (b) Contour of $\log p(\kappa_1, \kappa_2)$ (c) $y_i - \mathbb{E}[\beta_i|\mathbf{y}]$ v.s. y_i

Fig. 7. Set cover prior. The contour levels in panel (b) are: $-1, 0, 0.5, 1, 2$.



(a) Contour of $\mathbb{E}[\kappa_3|\kappa_1, \kappa_2]$ (b) Contour of $\log p(\kappa_1, \kappa_2)$ (c) $y_i - \mathbb{E}[\beta_i|\mathbf{y}]$ v.s. y_i

Fig. 8. Group counting prior. The contour levels in panel (b) are: $-2, -1, 0, 1, 2, 3$.



(a) Contour of $\mathbb{E}[\kappa_3|\kappa_1, \kappa_2]$ (b) Contour of $\log p(\kappa_1, \kappa_2)$ (c) $y_i - \mathbb{E}[\beta_i|\mathbf{y}]$ v.s. y_i

Fig. 9. MRF prior ($\alpha = 0.01$). The contour levels in panel (b) are -4, -1, -0.75, ..., 0, 1, 2.

Consider $p = 3$ variables, and there are two groups $\{1, 2\}$ and $\{2, 3\}$ which overlap on variable 2. The design matrix X is the 3×3 identity matrix I ($n = 3$), and the observation $\mathbf{y}|\boldsymbol{\beta} \sim \mathcal{N}(X\boldsymbol{\beta}, I) = \mathcal{N}(\boldsymbol{\beta}, I)$. Let $\sigma_0 = 1$. Then the expected posterior value of $\boldsymbol{\beta}$ given $\mathbf{z}$ has a closed form $\mathbb{E}[\beta_i|z_i, y_i] = (1-\kappa_i)y_i$ where $\kappa_i = 1/(1+z_i)$ is a

random shrinkage coefficient. The distribution of κ_i is determined entirely by the prior on z_i, and a larger value of κ_i means a greater amount of shrinkage towards the origin.

As a concrete example, we study the horseshoe prior with $a = b = d = 1/2$. GCP and SCP use the formulae (12) and (14), respectively. The MRF prior attaches a horseshoe potential on each β_i, and in addition employs a smooth MRF $\exp(-\alpha(z_1 - z_2)^2 - \alpha(z_2 - z_3)^2)$ with $\alpha = 0.01$. We use a Gaussian MRF because there is no need of shrinking the difference.

4.1 Conditional Dependency (Explain-away Effect)

We first consider the conditional distribution of κ_3 given κ_1 and κ_2. Since it is hard to visualize a function of three variables, we show in panel (a) of Figures 7 to 9 the mean $\mathbb{E}[\kappa_3|\kappa_2, \kappa_1]$ under the three priors. Clearly the mean of κ_3 does not change with κ_1 in GCP and MRF prior, because z_3 is simply independent of z_1 given z_2. The mean of κ_3 grows monotonically with κ_2, as MRF favors small difference between z_2 and z_3 (hence between κ_2 and κ_3), and in SCP smaller κ_2 clamps a larger value of $\max\{z_2, z_3\}$, shifting more probability mass towards greater z_3 which results in a lower mean of κ_3.

Interestingly when κ_2 is large, the SCP allows the mean of κ_3 to decrease when κ_1 grows. See, *e.g.*, the horizonal line at $\kappa_2 = 0.8$ in Figure 7a. To interpret this "explain-away" effect, first note a greater value of κ_2 means z_2 has a higher inverse scale. Due to the $\max$ in (14), it implies that either u_1 or u_2 is large, which means κ_1 or κ_3 is large since variables 1 and 3 belong to a single group only. Thus when κ_1 is small, κ_3 receives more incentive to be large, while this pressure is mitigated when κ_1 itself increases. On the other hand when κ_2 is small, κ_1 and κ_3 must be both small, leading to the flat contour lines.

4.2 Joint Shrinkage

Next we study the joint density of κ_1 and κ_2 plotted in panel (b). SCP exhibits a 2-D horseshoe shaped joint density in Figure 7b, which is desirable as it prefers either large shrinkage or little shrinkage. In GCP, however, the joint density of (κ_1, κ_2) concentrates around the origin in Figure 8b. Indeed, this issue arises even when there are only two variables, making a single group. Fixing κ_2 and hence z_2, $\max\{z_1, z_2\}$ does not approach 0 even when z_1 approaches 0 (*i.e.* κ_1 approaches 1). So it is unable to exploit the sparsity-inducing property of horseshoe prior which places a sharply growing density towards the origin. The joint density of MRF is low when κ_1 and κ_2 are both around the origin, although the marginal density of each of them seems still high around zero.

4.3 Robust Marginal Shrinkage

Finally we investigate the shrinkage profile via the posterior mean $\mathbb{E}[\boldsymbol{\beta}|\mathbf{y}]$, with $\mathbf{z}$ integrated out. Let $q(\mathbf{z})$ be proportional to the prior density on $\mathbf{z}$ (note the group counting and MRF priors need a normalizer). Then $\mathbb{E}[\beta_i|\mathbf{y}] = \gamma_i^{(1)}/\gamma_i^{(0)}$, where for $k \in \{0, 1\}$

$$\gamma_i^{(k)} = \int \beta_i^k q(\mathbf{z}) \prod_j \mathcal{N}(\beta_j; 0, z_j)\mathcal{N}(y_j; \beta_j, 1)\mathrm{d}\beta_j \mathrm{d}\mathbf{z}, \tag{15}$$

and $$\int \beta_j^k \mathcal{N}(\beta_j; 0, z_j)\mathcal{N}(y_j; \beta_j, 1)\mathrm{d}\beta_j = \sqrt{\frac{1}{8\pi}} \frac{z_j^k}{(1+z_j)^{k+\frac{1}{2}}} \exp\left(\frac{-y_j^2}{2+2z_j}\right). \tag{16}$$

Panel (c) of Figures 7 to 9 shows $y_i - \mathbb{E}[\beta_i|y_i]$ (the amount of shrinkage) as a function of y_i, for variables $i \in \{1, 2\}$. All y_j ($j \neq i$) are fixed to 1. In Figure 7c, Both SCP and GCP provide valuable robust shrinkage, with reasonable shrinkage when y_i is small in magnitude, and diminishes as y_i grows. And as expected, variable 2 shrinks more than variable 1. In SCP, variable 2 takes the sparser state between variables 1 and 3 via the max in (14), while in GCP variable 2 contributes to both sparsity-inducing priors of u_1 and u_2 in (12). Notice that for small y_1, GCP is not able to yield as much shrinkage as SCP. This is because for small y_1, z_1 is believed to be small, and hence the value of $\max\{z_1, z_2\}$ is dominated by the belief of z_2 (which is larger). This prevents z_1 from utilizing the horseshoe prior around zero. The case for y_2 is similar.

In fact, we can theoretically establish the robust shrinkage of SCP for any group structure under the current likelihood $\mathbf{y}|\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}, I)$.

Theorem 1. *Suppose SCP uses horseshoe prior in* (14) *with* $a = b = 1/2$. *Then for any group structure,* $\lim_{y_i \to \infty}(y_i - \mathbb{E}[\beta_i|\mathbf{y}]) = 0$ *with fixed values of* $\{y_j : j \neq i\}$.

Proof. (sketch) The key observation based on (15) and (16) is that $\mathbb{E}[\beta_i|\mathbf{y}] - y_i = \frac{\partial}{\partial y_i} \log F(\mathbf{y})$ where $F(\mathbf{y})$ is given by

$$\int_{\mathbf{z}} \prod_j (1+z_j)^{\frac{-1}{2}} \exp\left(\frac{-y_j^2}{2+2z_j}\right) q(\mathbf{z}) \mathrm{d}\mathbf{z} \tag{17}$$

$$= \int_{\mathbf{u}} \prod_j \left(\int_{z_j} (1+z_j)^{\frac{-1}{2}} \exp\left(\frac{-y_j^2}{2+2z_j}\right) \mathrm{Ga}(z_j; a, \max_{g:j\in g} u_g) \mathrm{d}z_j \right) \prod_g \mathrm{Ga}(u_g; b, d) \mathrm{d}\mathbf{u}$$

The rest of the proof is analogous to [6, Theorem 3]. The detailed proof is provided in Appendix B of the longer version of the paper [37]. □

By contrast, MRF is unable to drive down the amount of shrinkage when the response y_i is large. To see the reason (*e.g.* for variable 1), note we fix y_2 to 1. Since MRF enforces smoothness between z_1 and z_2, the fixed value of y_2 (hence its associated belief of z_2) will prevent z_1 to follow the increment of y_1, disallowing z_1 to utilize the heavy tail of horseshoe prior. The amount of shrinkage gets larger when α increases.

To summarize, among the three priors only the set cover prior enjoys all the three desirable properties namely conditional dependency, significant shrinkage for small observation, and vanishing shrinkage for large observations.

5 Experimental Results

We next study the empirical performance of SCP, compared with GCP, and MRF priors [34]. Since the MRF prior therein is restricted to spike and slab, to simplify comparison we also base SCP and GCP on spike and slab. This allows convenient application of expectation propagation for posterior inference [EP, 39, 40], where all discrete factors are approximated by Bernoulli messages [34]. At each iteration, messages are passed from top to the bottom in Figure 4, and back up. Other inference algorithms are also possible, such as MCMC [*e.g.*, 38], and variational Bayes [41]. Since the Bayesian models used

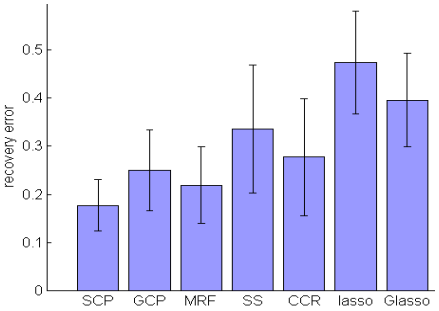


Fig. 10. Recovery rate for sparse signal

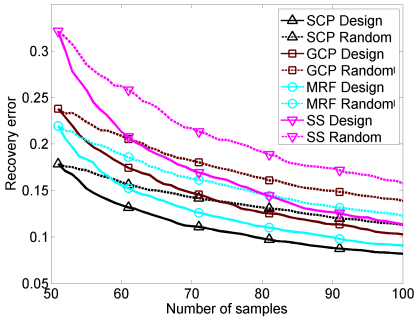


Fig. 11. Sequential experimental design for sparse recovery

here are typically multi-modal and the mean of the posterior is generally more important, we choose to use EP in our experiment, although it will also be interesting to try other methods.

Empirically EP always converged within 10 iterations, with change of message fallen below 1e-4. The loops make it hard to analyze the local or global optimality of EP result. But in practice, we did observe that with different initializations, EP always converged to the same result on all experiments, being highly reliable. To give an example of computational efficiency, in image denoising (Section 5.5, $p = n = 4096$), it took only 0.5 seconds per image and per iteration to compute messages related to the prior, while common techniques for Gaussian likelihood allowed its related messages to be computed in 1-2 seconds.

As a baseline, we also tried spike and slab prior with non-overlapping groups (GSS) if reasonable non-overlapping group approximation is available, or even without groups (SS). Furthermore we consider three state-of-the-art frequentist methods, including Lasso, group Lasso (GLasso), and coding complexity regularization [CCR, 26]. Groups are assumed available as prior knowledge.

5.1 Sparse Signal Recovery

We first consider a synthetic dataset for sparse signal reconstruction with $p = 82$ variables [42]. $\{\beta_i\}$ was covered by 10 groups of 10 variables, with an overlap of two variables between two successive groups: $\{1, \ldots, 10\}, \{9, \ldots, 18\}, \ldots, \{73, \ldots, 82\}$. The support of $\boldsymbol{\beta}$ was chosen to be the union of group 4 and 5, with the non-zero entries generated from *i.i.d.* Gaussian $\mathcal{N}(0, 1)$. We used $n = 50$ samples, with the elements of the design matrix $X \in \mathbb{R}^{n \times p}$ and the noisy measurements $\mathbf{y}$ drawn by

$$X_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1), \quad \mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1). \tag{18}$$

We used recovery error as the performance measure, which is defined as $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_2 / ||\boldsymbol{\beta}||_2$ for the posterior mean $\hat{\boldsymbol{\beta}}$. X and $\boldsymbol{\beta}$ were randomly generated for 100 times, and we report the mean and standard deviation of recovery error. An extra 10 runs were

Table 1. Recovery error for network sparsity. GSS is not included as disjoint group approximation is not clear for general graph structure. CCR is also not included since its implementation in [26] does not allow flexible specification of groups.

	SCP	GCP	MRF	SS	Lasso	GLasso
Jazz	**0.264** $\pm$ 0.083	0.312 $\pm$ 0.068	0.338 $\pm$ 0.149	0.398 $\pm$ 0.188	0.489 $\pm$ 0.101	0.456 $\pm$ 0.107
NetScience	**0.067** $\pm$ 0.005	0.093 $\pm$ 0.058	0.167 $\pm$ 0.110	0.188 $\pm$ 0.113	0.394 $\pm$ 0.045	0.383 $\pm$ 0.048
Email	**0.106** $\pm$ 0.025	**0.104** $\pm$ 0.054	0.243 $\pm$ 0.105	0.310 $\pm$ 0.130	0.432 $\pm$ 0.049	0.420 $\pm$ 0.057
C.elegans	**0.158** $\pm$ 0.034	0.163 $\pm$ 0.025	0.184 $\pm$ 0.057	0.225 $\pm$ 0.101	0.408 $\pm$ 0.068	0.394 $\pm$ 0.068

taken to allow all models to select the hyper-parameters that optimize the performance on the 10 runs. This scheme is also used in subsequent experiments.

In Figure 10, SCP clearly achieves significantly lower recovery error than all other methods. MRF is the second best, followed by GCP. This suggests that when $\boldsymbol{\beta}$ is generated over the union of some groups, SCP is indeed most effective in harnessing this knowledge. Bayesian models for structured sparse estimation also outperform vanilla Bayesian models for independent variables (SS), as well as frequentist methods (CCR, Lasso, GLasso),

5.2 Sequential Experimental Design

A key advantage of Bayesian model is the availability of uncertainty estimation which facilitates efficient sequential experimental design [13]. We randomly generated a data pool of 10,000 examples based on (18), and initialized the training set with $n = 50$ randomly selected examples (*i.e.* revealing their response y_i). Then we gradually increased the size of training set up to $n = 100$. At each iteration, one example was selected and its response y_i was revealed for training. In the random setting examples were selected uniformly at random, while in sequential experimental design, typically the example with the highest uncertainty was selected. For each candidate example $\mathbf{x}$, we used $\mathbf{x}'V\mathbf{x}$ as the uncertainty measure, where V is the current approximated posterior covariance matrix. The whole experiment was again repeated for 100 times, and the average recovery error is shown.

In Figure 11, for all models sequential experimental design is significantly more efficient in reducing the recovery error compared with random design. In particular, SCP achieves the steepest descent in error with respect to the number of measurements. This again confirms the superiority of SCP in modeling group structured sparsity in comparison to GCP and MRF. SS performs worst as it completely ignores the structure.

5.3 Network Sparsity

Following [34] and [43], we next investigate the network sparsity where each node is a variable and each edge constitutes a group (*i.e.* all groups have size 2). We tried on four network structures: Email ($p = 1,133$, #edge=5,451), C.elegans (453, 2,015), Jazz (198, 2,742), NetScience (1,589, 2,742).[3] See network properties in Table 1. We picked a subset of edges uniformly at random, and added their two incident nodes to the support of $\boldsymbol{\beta}$. By adjusting the number of selected edges, the size of the support of $\boldsymbol{\beta}$ is

[3] Downloaded from http://www-personal.umich.edu/∼mejn/netdata

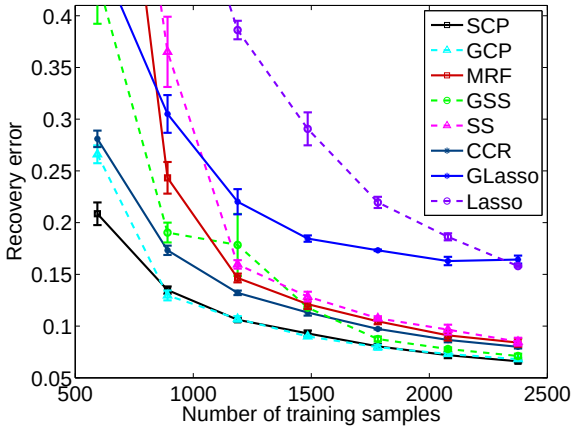


Fig. 12. Recovery error for background subtraction

$0.25p$, and the nonzero elements in $\boldsymbol{\beta}$ were sampled from $\mathcal{N}(0,1)$. The design matrix X and the response $\mathbf{y}$ were drawn from (18). We used $n = \lfloor p/2 \rfloor$ examples as in [43].

The average recovery error of 100 runs is shown in Table 1. Again SCP yields significantly lower error than all other algorithms, except for a tie with GCP on Email. GCP outperforms MRF, which in turn defeats all other methods that do not faithfully model the group structure.

5.4 Background Subtraction

We next consider real-world applications in compressive sensing with overlapped group sparsity. Here the data generating process is beyond our control. In video surveillance, the typical configuration of the images are the sparse foreground objects on the static backgrounds. Our task here is to recover the sparse background subtracted images via compressive sensing.

Our experimental setting follows [26][4]. The spatial consistency is an important prior knowledge on 2D image signals which has been successfully leveraged in various applications. Specifically pixels in a spatial neighborhood are likely to be background or foreground at the same time. Edges connecting pixels to its four neighbors are used in the MRF prior to encourage the consistency between adjacent pixels. For GSS which requires no overlap between groups, we simply defined the groups as non-overlapped 3×3 patches. For the rest structured priors, we defined groups as the overlapped 3×3 patches. Singleton groups were also added to deal with isolated foreground pixels. Each image is sized 80×80 ($p = 6,400$). We varied the number of image (n) from 600 to 2400.

Figure 12 shows SCP and GCP achieve significantly lower recovery error than other methods on any number of measurement. The prior of spatial consistency does help improve the recovery accuracy, especially when the size of the training set is small. With sufficient training samples, both structured and non-structured methods can have accurate recovery. This can be seen by comparing Lasso with GLasso, as well as SCP

[4] Video from http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1

Table 2. PSNR in image denoising. MRF and GSS are not included because in the case of hierarchical structure, it is not clear how to enforce MRF, or approximate a tree by disjoint groups.

	SCP	GCP	SS	Lasso	GLasso	CCR
House	**28.77** $\pm$ 0.04	28.13 $\pm$ 0.06	27.72 $\pm$ 0.06	27.22 $\pm$ 0.02	27.24 $\pm$ 0.03	27.79 $\pm$ 0.04
Lenna	**28.27** $\pm$ 0.03	27.65 $\pm$ 0.02	27.28 $\pm$ 0.02	26.95 $\pm$ 0.03	27.15 $\pm$ 0.01	27.11 $\pm$ 0.02
pepperr	**26.57** $\pm$ 0.03	25.87 $\pm$ 0.01	25.75 $\pm$ 0.03	25.06 $\pm$ 0.05	25.39 $\pm$ 0.06	25.51 $\pm$ 0.04
Boat	**26.80** $\pm$ 0.01	26.24 $\pm$ 0.01	26.09 $\pm$ 0.01	25.65 $\pm$ 0.01	26.05 $\pm$ 0.02	25.65 $\pm$ 0.01
Barbara	**24.93** $\pm$ 0.02	24.56 $\pm$ 0.02	24.43 $\pm$ 0.02	24.23 $\pm$ 0.01	24.77 $\pm$ 0.02	24.34 $\pm$ 0.01

with GCP, GSS, and SS. The superiority of SCP and GCP over GSS corroborates the importance of accommodating more flexible group definitions.

5.5 Image Denoising with Tree-structured Wavelets

Our last set of experiment examines the effectiveness of structured sparse priors for modeling hierarchical sparsity. The task is to restore 2D images which are contaminated by noise via compressive sensing on 2D wavelet basis. The setting is similar to [26] and [21]. 2D wavelet basis at different resolution levels is used as dictionary to get sparse representation of images. There is a natural hierarchical structure in the wavelet coefficients: a basis $\mathbf{b}$ can be defined as the parent of all such basis at finer resolution and whose support is covered by the support of $\mathbf{b}$. Such tree-structured dependency corresponds to the nature of multi-resolution wavelet analysis and have been proven empirically effective in sparse representation of signals.

We choose the orthogonal Haar wavelet basis and the classical quad-tree structure on the 2D wavelet coefficients. We use PSNR $:= \log_{10}(\frac{255^2}{\text{MSE}})$ to measure the quality of recovery. The benchmark set consists of five standard testing images: house, Lenna, boat, Barbara and pepper. We added Gaussian white noise $\mathcal{N}(0, 25^2)$ to the original images. The PSNR of the resulting noisy image is around 20. The images were divided into non-overlapped patches sized 64×64. Each patch is recovered independently with six levels of 2D Haar wavelet basis. For each method, we selected the parameters with the highest PSNR.

The recovery result is shown in Table 2. SCP delivers the highest PSNR in denoising on all test images, demonstrating the power of hierarchical structure prior to improve the recovery accuracy. Figure 15 in Appendix C of [37] shows a visual comparison of the denoising results, and it can be observed that SCP outperforms other methods in removing noise and preserving details in the image.

6 Conclusion and Discussion

We proposed a framework of set cover prior for modeling structured sparsity with overlapped groups. Its behavior is analyzed and empirically it outperforms existing competent structured priors. For future work, it will be interesting to further model sparsity within each group [18, 44]. Extension to other learning tasks is also useful, *e.g.* multi-task learning [45, 46].

Acknowledgements NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

[1] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data*. Springer, 2011.

[2] Y. Eldar and G. Kutyniok, editors. *Compressed Sensing: Theory and Applications*. Cambridge, 2012.

[3] E. George and R. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88:881–889, 1993.

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

[5] A. Armagan, D. Dunson, and M. Clyde. Generalized Beta mixtures of Gaussians. In *NIPS*, 2011.

[6] C. Carvalho, N. Polson, and J. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97:465–480, 2010.

[7] C. Carvalho, N. Polson, and J. Scott. Handling sparsity via the horseshoe. In *AI-STATS*, 2009.

[8] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of The Royal Statistical Society, Series B*, 58:267–288, 1996.

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

[10] T. Park and G. Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):618–686, 2008.

[11] J. Griffin and P. Brown. Bayesian adaptive lassos with non-convex penalization. *Australian & New Zealand Journal of Statistics*, 53(4):423–442, 2011.

[12] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.

[13] D. Hernández-Lobato, J. M. Hernández-Lobato, and P. Dupont. Generalized spike and slab priors for Bayesian group feature selection using expectation propagation. *Journal of Machine Learning Research*, 16:1891–1945, 2013.

[14] S. Ji, D. Dunson, and L. Carin. Multitask compressive sensing. *IEEE Trans. Signal Processing*, 57(1):92–106, 2009.

[15] S. Raman, T. Fuchs, P. Wild, E. Dahl, and V. Roth. The Bayesian group-lasso for analyzing contingency tables. In *ICML*, 2009.

[16] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society, Series B*, 68(1):49–67, 2006.

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

[18] T.-J. Yen and Y.-M. Yen. Grouped variable selection via nested spike and slab priors. arXiv 1106.5837, 2011.

[19] Y. Suo, M. Dao, T. Tran, U. Srinivas, and V. Monga. Hierarchical sparse modeling using spike and slab priors. In *ICASSP*, 2013.

[20] C. Li and H. Li. Network-constrained regularization and variable selection for analysis of genomic data. *Biometrics*, 24(9):1175–1182, 2008.

[21] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for hierarchical sparse coding. *Journal of Machine Learning Research*, 12:2297–2334, 2011.

[22] F. Li and N. Zhang. Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association*, 105 (491):1201–1214, 2010.

[23] W. Pan, B. Xie, and X. Shen. Incorporating predictor network in penalized regression with application to microarray data. *Biometrics*, 66(2):474–484, 2010.

[24] F. Stingo and M. Vannucci. Variable selection for discriminant analysis with Markov random field priors for the analysis of microarray data. *Bioinformatics*, 27(4):495–501, 2011.

[25] P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. *Annals of Stat.*, 37(6A):3468–3497, 2009.

[26] J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. *Journal of Machine Learning Research*, 12:3371–3412, 2011.

[27] J. Griffin and P. Brown. Hierarchical sparsity priors for regression models. arXiv:1307.5231, 2013.

[28] M. Yuan, V. R. Joseph, and H. Zou. Structured variable selection and estimation. *Annals of Applied Statistics*, 3:1738–1757, 2009.

[29] J. Griffin and P. Brown. Some priors for sparse regression modelling. *Bayesian Analysis*, 8 (3):691–702, 2013.

[30] J. A. Palmer, D. P.Wipf, K. Kreutz-Delgado, and B. D. Rao. Variational EM algorithms for non-Gaussian latent variable models. In *NIPS*, 2005.

[31] J. Bernardo and A. Smith. *Bayesian Theory*. Wiley, 1994.

[32] J. Griffin and P. Brown. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188, 2010.

[33] G. Obozinski and F. Bach. Convex relaxation for combinatorial penalties. Technical Report HAL 00694765, 2012.

[34] J. M. Hernández-Lobato, D. Hernández-Lobato, and A. Suárez. Network-based sparse Bayesian classification. *Pattern Recognition*, 44(4):886–900, 2011.

[35] A. Argyriou, R. Foygel, and N. Srebro. Sparse prediction with the k-support norm. In *NIPS*, 2012.

[36] L. He and L. Carin. Exploiting structure in wavelet-based Bayesian compressive sensing. *IEEE Trans. Signal Processing*, 57(9):3488–3497, 2009.

[37] X. Liu, X. Zhang, and T. Caetano. Bayesian models for structured sparse estimation via set cover prior. Technical report, 2014. http://users.cecs.anu.edu.au/~xzhang/papers/LiuZhaCae14_long.pdf.

[38] F. Stingo, Y. Chen, M. Tadesse, and M. Vannucci. Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes. *Annals of Applied Statistics*, 5(3):1978–2002, 2011.

[39] T. Minka. *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, MIT, 2001.

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

[41] P. Carbonetto and M. Stephens. Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis*, 7(1): 73–108, 2012.

[42] L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *ICML*, 2009.

[43] J. Mairal and B. Yu. Supervised feature selection in graphs with path coding penalties and network flows. *Journal of Machine Learning Research*, 14:2449–2485, 2013.

[44] V. Rockova and E. Lesaffre. Incorporating grouping information in Bayesian variable selection with applications in genomics. *Bayesian Analysis*, 9(1):221–258, 2014.

[45] D. Hernández-Lobato and J. M. Hernández-Lobato. Learning feature selection dependencies in multi-task learning. In *NIPS*, 2013.

[46] D. Hernández-Lobato, J. M. Hernández-Lobato, T. Helle-putte, and P. Dupont. Expectation propagation for Bayesian multi-task feature selection. In *ECML*, 2010.