Learning to Learn for Structured Sparsity

Nino Shervashidze
INRIA - Sierra Project-Team
École Normale Supérieure
Paris, France
nino.shervashidze@inria.fr

Francis Bach
INRIA - Sierra Project-Team
École Normale Supérieure
Paris, France
francis.bach@inria.fr

May 2, 2014

Abstract

Structured sparsity has recently emerged in statistics, machine learning and signal processing as a promising paradigm for learning in high-dimensional settings. A number of methods have been proposed for learning under the assumption of structured sparsity, including group LASSO and graph LASSO. All of these methods rely on prior knowledge on how to weight (equivalently, how to penalize) individual subsets of variables during the subset selection process. However, these weights on groups of variables are in general unknown. Inferring group weights from data is a key open problem in research on structured sparsity.

In this paper, we propose a Bayesian approach to the problem of group weight learning. We model the group weights as hyperparameters of heavy-tailed priors on groups of variables and derive an approximate inference scheme to infer these hyperparameters. We empirically show that we are able to recover the model hyperparameters when the data are generated from the model, and moreover, we demonstrate the utility of learning group weights in synthetic and real denoising problems.

1 Introduction

High-dimensional prediction problems are more and more common in many application domains such as computational biology, computer vision, signal processing or natural language processing. To handle this high-dimensionality, one usually resorts to linear modeling and regularization with sparsity-inducing norms, such as the ℓ_1 norm. This type of regularization results in *sparse* models, meaning that the model is described by relatively few parameters. Besides making parameter inference easier, the sparsity assumption has the appealing property of yielding more interpretable models. As an example, consider the problem of explaining a particular phenotype of patients, e.g., the disease state, based on the genome sequence of each patient. Sparse linear approaches try to find a handful of genome loci that govern the disease state, rather than a model involving the whole sequence. The ℓ_1-regularized sparse linear models, such as the LASSO (Tibshirani, 1994), are well studied by now, with a solid body of theoretical results, efficient algorithms and applications in diverse fields (see, e.g., Bühlmann & van de Geer, 2011, and references therein). However, in practice, we often know that there is more *structure* in the problem at hand, which cannot be captured by simple sparse modeling and ℓ_1 regularization, and which, if exploited, can improve the estimation of parameters as well as the interpretability of the estimates (see Cevher et al., 2008; Huang et al., 2011; Bach et al., 2012b, and references therein). In our example, we could expect the genetic loci that influence the disease to be part of a small number of connected patterns in a known gene-gene interaction network (e.g., Rapaport et al., 2007; Azencott et al., 2013). In other words, we are looking for a small number of possibly overlapping subsets of variables such that each subset corresponds to a connected subgraph in the given gene network, and the combination of variables in each subset influences the phenotype.

Given prior knowledge about the relevance of each considered group of variables, several methods exist for learning sparse models guided by this prior knowledge. These methods achieve

different kinds of structured sparsity by regularization (penalization, weighting) with appropriate sparsity-inducing norms, that often correspond to convex relaxations of combinatorial penalties on the support (i.e., non-zero pattern) of the parameter vector. After the group LASSO (Yuan & Lin, 2006), a number of convex penalties have been proposed, generalizing the group LASSO penalty to the cases of overlapping groups (e.g., Zhao et al., 2009; Jacob et al., 2009; Jenatton et al., 2011a), including tree-structured groups (Kim & Xing, 2010; Jenatton et al., 2011b). See the monograph by Bach et al. (2012a, Section 1.3) for a more detailed review of sparsity-inducing norms.

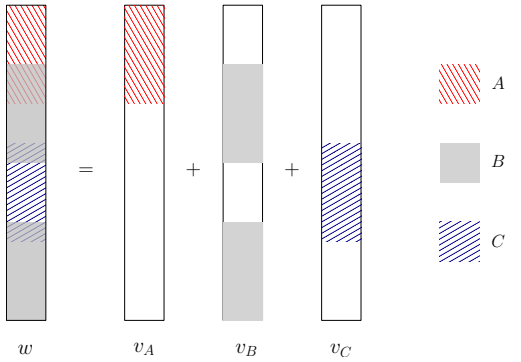

Figure 1: The coefficient vector w is covered by latent variables supported on subsets A, B and C: $w = v_A + v_B + v_C$.

While most of these norms induce *intersection-closed* sets of non-zero patterns, Jacob et al. (2009) and Obozinski & Bach (2012) introduce a different, latent formulation of sparsity-inducing norms that yields *union-closed* sets of non-zero patterns, meaning that the parameter vector w is represented as a sum of latent vectors v_A, identically zero at indices not in A (see Figure 1 for illustration). Moreover, Obozinski & Bach (2012) show that the sparsity-inducing penalties g on the support $\mathrm{supp}(w)$ of w that are adapted to convex relaxations can be written as

$$g(\mathrm{supp}(w)) = \min_{\substack{\mathcal{A}' \subseteq \mathcal{A}, \\ \cup_{A \in \mathcal{A}'} A = \mathrm{supp}(w)}} \sum_{A \in \mathcal{A}'} f(A), \quad (1)$$

that is, $g(\mathrm{supp}(w))$ is the minimum-weight *cover* of $\mathrm{supp}(w)$ with the subsets A in the family $\mathcal{A}$. The weights $f(A)$ express our prior belief in the subset A being relevant: If a group A is irrelevant, then $f(A) = \infty$. The corresponding norm is then

$$\Omega(w) = \min_{\substack{v_A \in \mathbb{R}^P, \\ \sum_{A \in \mathcal{A}} v_A = w}} \sum_{A \in \mathcal{A}} \|v_A\|_2 f(A)^{1/2}. \quad (2)$$

However, generally we do not have this prior knowledge about the relevance of individual groups: The problem of automatically choosing appropriate weights for groups of variables, $f(A)$, is a key open problem in research on structured sparsity. Assuming that we have several learning problems with similar structure (the relevance of a given group is largely shared across individual problems), in this paper we propose a framework for learning group relevances from data. Note that learning the structure is naturally a multi-task problem, as it is impossible to estimate the prior on a vector of parameters if we only observe one particular instance of it. To come back to our example, we could assume that we have several phenotypes that can be explained by groups of loci whose relevance is largely shared across phenotypes. A recent approach to learning group relevances from data has been proposed by Hernández-Lobato & Hernández-Lobato (2013). However, this work only considers learning relevances of pairs of variables and does not make the link with sparsity-inducing norms.

We approach the problem using probabilistic modeling with a broad family of heavy-tailed priors and derive a variational inference scheme to learn the parameters of these priors. Our model follows the pattern of *sparse Bayesian* models (Palmer et al., 2006; Seeger & Nickisch, 2011, among others), that we take two steps further: First, we propose a more general formulation, suitable for structured sparsity with any family of groups; Second, we learn from data the prior parameters that are supposed to be given and most often assumed to be common to all variables in existing work. We show that prior parameter estimation with classical variational inference does not always lead to reasonable estimates in these models, and find a way of regularizing that works well in practice. In our experiments, we show that we are able to recover the model parameters when the data are generated from the model, and moreover, we demonstrate the utility of learning group penalties in image denoising.

2 A Probabilistic Model for Structured Sparse Linear Regression

In this section we formally describe our model and a suitable approximate inference scheme for this model.

2.1 Model definition

We consider K linear regression problems with design matrices $X^k \in \mathbb{R}^{N\times P}$ and response vectors $y^k \in \mathbb{R}^N$ for $k \in \{1,\dots,K\}$. For each X^k and y^k, we assume the classical Gaussian linear model with i.i.d. noise with variance σ^2, that is,

$$y^k \sim \mathcal{N}(X^k w^k, \sigma^2 I).$$

Let V be the set of indices of variables $\{1,\dots,P\}$. For a family $\mathcal{A}$ of subsets of V, we assume

$$w^k = \sum_{A\in\mathcal{A}} v_A^k, \tag{3}$$

where, for each k,

- $\forall A \in \mathcal{A}, v_A^k$ is a vector in $\mathbb{R}^P$ such that all its components with indices in $V \setminus A$ are zero (in other words, it is supported on A),
- $\{v_A^k\}_{A\in\mathcal{A}}$ are jointly independent, and
- $\forall A \in \mathcal{A}, v_A^k$ has an isotropic density with inverse scale parameter $f(A)$

$$p(v_A^k|f(A)) = q_A(\|v_A^k\|_2 f(A)^{1/2}) f(A)^{|A|/2}, \tag{4}$$

where q_A is a heavy-tailed distribution that only depends on A through its cardinality, $|A|$. We specify q_A in Section 2.2.

Finally, as v_A^k are assumed independent,

$$p(w^k|f) = \prod_{A\in\mathcal{A}} p(v_A^k|f(A)). \tag{5}$$

We regard the inverse scale parameter $f(A)$ as a measure of relevance of the group of variables A[1]: If a group of variables is irrelevant, then $f(A)$ should equal infinity. We are interested in priors q_A such that for each task indexed by k only a handful of v_A^k can be significantly away from zero.

Here it is important to stress the link between the expression of our isotropic prior (4) and the norm (2) by Obozinski & Bach (2012): The log-likelihood of parameter vectors $\{w^k\}_{k=1,\dots,K}$ with respect to f will (up to a constant) be equal to the term $\sum_{A\in\mathcal{A}} \log q_A(\|v_A^k\|_2 f(A)^{1/2})$, which very closely resembles the norm (2). If q_A is the generalized Gaussian distribution (cf. Section 2.4), the two expressions match exactly. Thus, learning with our prior is a natural probabilistic counterpart of learning with the sparsity-inducing norm (2).

Given data $\{X^k, y^k\}_{k=1,\dots,K}$ and such a model for the prior, our goal will be to infer the parameters $f(A)$ by maximizing the likelihood with respect to f,

$$p(y^1,\dots,y^K|f) = \prod_{k=1}^K \int p(y^k|X^k w^k, \sigma^2 I) \prod_{A\in\mathcal{A}} p(v_A^k|f(A)) dv_A^k, \tag{6}$$

where the parameters v_A^k are marginalized.

[1] Abusing notation, we will call "group A" the subset of variables indexed by elements of A throughout the paper.

2.2 Super-Gaussian priors

We assume that q_A is a *scale mixture of Gaussians*, i.e.,

$$q_A(u) = \int_0^\infty \mathcal{N}(u|0, s) r_A(s) ds$$

for some mixing density $r_A(s)$. The main reason why we choose to work with the family of scale mixtures of zero-mean Gaussians is that it contains distributions that are heavy-tailed and therefore suitable for modeling sparsity; One such distribution is Student's t that we use in our experiments. The inverse scale parameter of the distribution on v_A^k, $f(A)$, captures the relevance of the group A: the smaller $f(A)$, the more relevant the group, that is, the larger the values v_A^k is likely to take. Note that even if the group A is relevant, not all $v_A^k, k = 1, \ldots, K$ have to be large. In fact, if the parameters $v_A^k, k = 1, \ldots, K$ are drawn from a heavy-tailed distribution with small $f(A)$, then only a fraction of them will be significantly away from zero. Moreover, as we show in Section 2.3, learning in such models is amenable to variational optimization with closed-form updates and leads to an approximate Gaussian posterior on v_A^k.

In general, the integral in (6) is intractable for Gaussian scale mixtures, therefore one has to resort to sampling or approximate inference to learn parameters in such models. The fact that q_A is a Gaussian scale mixture implies that it is also *super-Gaussian*, that is, the logarithm of $q_A(u)$ is convex in u^2 and non-increasing (Palmer et al., 2006). It therefore admits a representation of the following form by convex conjugacy

$$\log q_A(u) = \sup_{s \geq 0} -\frac{u^2}{2s} - \phi_A(s), \tag{7}$$

where $\phi_A(s)$ is convex in $1/s$. Note that the expression inside the supremum in (7) has a unique maximizer that we denote $s^*(u)$. From (7), we get the following variational representation for $p(v_A^k|f(A))$:

$$\begin{aligned} p(v_A^k|f(A)) &= f(A)^{\frac{|A|}{2}} \sup_{\zeta_A^k \geq 0} \exp\Big(-\frac{\|v_A^k\|_2^2 f(A)}{2\zeta_A^k} - \phi_A(\zeta_A^k)\Big) \\ &= f(A)^{\frac{|A|}{2}} \sup_{\zeta_A^k \geq 0} \Big[\mathcal{N}\Big(v_A^k\Big|0, \frac{\zeta_A^k I}{f(A)}\Big)\Big(2\pi\frac{\zeta_A^k}{f(A)}\Big)^{\frac{|A|}{2}} e^{-\phi_A(\zeta_A^k)}\Big]. \end{aligned}$$

For a particular choice of the prior q_A, we measure the relevance of the group of variables A by the expectation of $\|v_A^k\|_2^2$ (which amounts to the sum of the variances of the individual components of v_A^k),

$$\mathbb{E}\left[\|v_A^k\|_2^2\right] = \frac{\mathbb{E}_{\|z\|_2 \sim q_A}\left[\|z\|_2^2\right]}{f(A)}, \tag{8}$$

where $\mathbb{E}_{\|z\|_2 \sim q_A}\left[\|z\|_2^2\right]$ is the expectation of $\|z\|_2^2$ under the standardized distribution q_A on $\|z\|_2$. In fact, as we have

$$\mathbb{E}\left[\|w^k\|_2^2\right] = \sum_{A \in \mathcal{A}} \mathbb{E}\left[\|v_A^k\|_2^2\right] \tag{9}$$

given our independence assumption, the expected value of $\|v_A^k\|_2^2$ allows us to measure the contribution of the group A with respect to $\mathbb{E}\left[\|w^k\|_2^2\right]$. We somewhat abusively call $\mathbb{E}\left[\|w^k\|_2^2\right]$ the *signal variance* in our experiments, as opposed to $P\sigma^2$, the *noise variance*. Figure 2 represents the graphical model corresponding to our assumptions. Note that we have explicitly incorporated the variational parameter ζ_A^k into the graphical model: In fact, the same parameter can also be interpreted as the scale parameter of the Gaussian in the Gaussian scale mixture representation of $p(v_A^k|f(A))$ (Palmer et al., 2006).

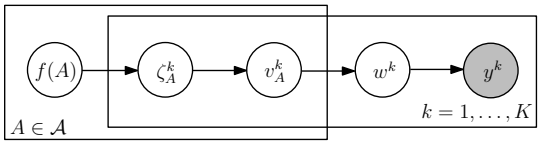

Figure 2: The graphical representation of our model.

2.3 Inference

The model described above leads to the following variational bound on the marginal distribution of y^k:

$$p(y^k|f) \leq \sup_{\substack{\zeta_A^k \geq 0 \\ A \in \mathcal{A}}} \Big\{ \log \mathcal{N}(y^k|0, X^k M Z^k F^{-1} M^\top {X^k}^\top + \sigma^2 I) + \sum_{A \in \mathcal{A}} \Big[\frac{|A|}{2} \log f(A) + \frac{|A|}{2} \log \Big(2\pi \frac{\zeta_A^k}{f(A)} \Big) - \phi_A(\zeta_A^k) \Big] \Big\},$$

where M is a matrix of dimension $P \times \sum_{A \in \mathcal{A}} |A|$ that ensures $w^k = M v^k$ where v^k is the concatenation of all elements indexed by elements of A in $v_A^k, A \in \mathcal{A}$, and F and Z^k are square diagonal matrices of size $\sum_{A \in \mathcal{A}} |A|$ whose diagonals consist of $f(A)$ and ζ_A^k respectively, replicated $|A|$ times, for each $A \in \mathcal{A}$. Thus, as an approximation to minimizing the negative log-likelihood, we would like to minimize the following overall bound with respect to f and ζ_A^k for all $A \in \mathcal{A}$ and $k \in \{1, \ldots, K\}$:

$$\begin{aligned} -\sum_{k=1}^{K} \Big\{ -\frac{1}{2} {y^k}^\top \Big(X^k M Z^k F^{-1} M^\top {X^k}^\top + \sigma^2 I \Big)^{-1} y^k - \frac{1}{2} \log\det \Big(X^k M Z^k F^{-1} M^\top {X^k}^\top + \sigma^2 I \Big) \\ + \sum_{A \in \mathcal{A}} \frac{|A|}{2} \log f(A) + \frac{\sum_{A \in \mathcal{A}} |A| - N}{2} \log(2\pi) + \frac{1}{2} \log\det(Z^k F^{-1}) - \sum_{A \in \mathcal{A}} \phi_A(\zeta_A^k) \Big\}. \end{aligned} \tag{10}$$

In its form given by (10), the bound is difficult to optimize. However, we recognize parts of it as minima of convex functions, which allows us to design an iterative algorithm with analytic updates, finding a local minimum (see the appendix for details). Our optimization problem becomes

$$\begin{aligned} \inf_{\zeta^k \geq 0} \inf_{v^k} \inf_{\Sigma^k \succcurlyeq 0} \sum_{k=1}^{K} \Big\{ \frac{1}{2\sigma^2} \|y^k - X^k M v^k\|_2^2 + \frac{1}{2} \sum_{A \in \mathcal{A}} \frac{f(A)}{\zeta_A^k} (\|v_A^k\|_2^2 + \operatorname{Tr} \Sigma_{AA}^k) - \frac{1}{2} \log\det \Sigma^k \\ + \frac{N}{2} \log(\sigma^2) + \frac{N}{2} \log(2\pi) + \frac{1}{2\sigma^2} \operatorname{Tr} M^\top {X^k}^\top X^k M \Sigma^k - \frac{1}{2} \sum_{A \in \mathcal{A}} |A| \\ + \sum_{A \in \mathcal{A}} \Big[-\frac{1}{2} |A| \log f(A) - \frac{|A|}{2} \log 2\pi + \phi_A(\zeta_A^k) \Big] \Big\}, \end{aligned} \tag{11}$$

and the closed-form updates are

$$\begin{aligned} \Sigma^k &= \sigma^2 (M^\top {X^k}^\top X^k M + \sigma^2 F {Z^k}^{-1})^{-1} \\ v^k &= (M^\top {X^k}^\top X^k M + \sigma^2 F {Z^k}^{-1})^{-1} M^\top {X^k}^\top y^k \\ \zeta_A^k &= \operatorname*{argmin}_{z \geq 0} \phi_A(z) + \frac{1}{2} \frac{f(A)}{z} (\|v_A^k\|_2^2 + \operatorname{Tr} \Sigma_{AA}^k) \\ \sigma^2 &= \frac{\sum_{k=1}^{K} \big\{ \|y^k - X^k M v^k\|_2^2 + \operatorname{Tr} M^\top {X^k}^\top X^k M \Sigma^k \big\}}{NK} \\ f(A) &= \frac{K|A|}{\sum_{k=1}^{K} \frac{1}{\zeta_A^k} (\|v_A^k\|_2^2 + \operatorname{Tr} \Sigma_{AA}^k)}, \end{aligned} \tag{12}$$

iterated until convergence.

Remark 2.1 *Note that the only update that depends on the specific prior distribution is that for the variational parameter ζ_A^k, all others apply to all super-Gaussian priors.*

Remark 2.2 *It can be shown that the updates* (12) *exactly correspond to the updates yielded by mean-field variational inference in the special case of Gaussian scale mixtures (Palmer et al., 2006). However, the approach presented here is more general, as it also applies to super-Gaussian priors that are not Gaussian scale mixtures.*

Remark 2.3 *Using matrix inversion lemma, the update for Σ^k can be rewritten in such a way that we avoid the expensive inversion of a $\sum_{A\in\mathcal{A}}|A| \times \sum_{A\in\mathcal{A}}|A|$ matrix and we only have to invert a $P \times P$ or $N \times N$ matrix instead (see the appendix for details). Note moreover that matrix inversions could be avoided altogether by making an extra diagonal assumption on the covariance matrix of the Gaussian posteriors of all v_A^k.*

Remark 2.4 *While we do provide an update equation for σ^2 for completeness, in general it is customary to assume the noise level known, which we also do in all our experiments.*

2.4 Special cases

The family of super-Gaussian distributions includes Student's t and generalized Gaussian distributions among many others. We here give the densities of these distributions, as well as the expressions for the quantities in our model and inference that depend on the particular prior on v_A^k.

Student's t: The density of this distribution is given by

$$p(v_A^k|a, f(A)) = f(A)^{\frac{|A|}{2}} \frac{\Gamma(a+|A|/2)}{\Gamma(a)} \Big(\frac{1}{2\pi}\Big)^{\frac{|A|}{2}} \Big(1 + \frac{\|v_A^k\|_2{}^2 f(A)}{2}\Big)^{-a-\frac{|A|}{2}}, \tag{13}$$

where a is a parameter governing the shape of the distribution. The smaller a, the heavier-tailed the distribution (for $a \leq 1$, there is no finite variance). For this distribution,

$$\begin{aligned}\phi_A(\zeta_A^k) =& \frac{1}{\zeta_A^k} + (a+1/2)\log(\zeta_A^k) + \frac{|A|}{2}\log(2\pi) - (a+|A|/2) + (a+|A|/2)\log(a+|A|/2) \\ & - \log(\Gamma(a+|A|/2)) + \log(\Gamma(a)),\end{aligned} \tag{14}$$

and, therefore, the update for ζ_A^k is written as

$$\zeta_A^k = \frac{1 + \frac{1}{2}f(A)(\|v_A^k\|_2^2 + \operatorname{Tr}\Sigma_{AA}^k)}{a + \frac{|A|}{2}}. \tag{15}$$

The variance of a Student's t-distributed random variable, if $a > 1$, is $\mathbb{E}(v_A^k {v_A^k}^\top) = \frac{1}{f(A)(a-1)} I$, and therefore $\mathbb{E}(\|v_A^k\|_2^2) = \frac{|A|}{f(A)(a-1)}$. Student's t has a natural representation as a Gaussian scale mixture with the inverse Gamma as the mixing distribution. All our experiments are carried out using Student's t.

Generalized Gaussian: The density is given by

$$p(v_A^k|\gamma, f(A)) = f(A)^{\frac{|A|}{2}} \frac{\frac{\gamma}{2}\Gamma(\frac{|A|}{2})}{\pi^{\frac{|A|}{2}}\Gamma(\frac{|A|}{2})} e^{-\|v_A^k f(A)^{\frac{1}{2}}\|_2^\gamma} \tag{16}$$

(Pascal et al., 2013). Consequently, we have

$$\phi_A(\zeta_A^k) = -\log\frac{\frac{\gamma}{2}\Gamma(\frac{|A|}{2})}{\pi^{\frac{|A|}{2}}\Gamma(\frac{|A|}{2})} + \frac{{\zeta_A^k}^{\frac{\gamma}{2-\gamma}}(\frac{1}{\gamma}-\frac{1}{2})}{\gamma^{\frac{2}{\gamma-2}}}, \tag{17}$$

$$\zeta_A^k = \Big(-\frac{\frac{1}{2}f(A)(\|v_A^k\|_2^2 + \operatorname{Tr}\Sigma_{AA}^k)}{(1/\gamma - 1/2)\gamma^{\frac{\gamma-2}{2}}}\frac{\gamma-2}{\gamma}\Big)^{\frac{2-\gamma}{2}}, \tag{18}$$

and $\mathbb{E}(\|v_A^k\|_2^2) = \frac{\Gamma(|A|/\gamma+2/\gamma)}{f(A)\Gamma(|A|/\gamma)}$.

3 Approximation Quality and Regularization

The goal of this section is to experimentally study the behavior of our approximate inference scheme in terms of estimation quality and to clarify how we can control it. As we empirically show below, the variational approximation scheme from Section 2.3 tends to overestimate the variance of the prior distribution (i.e., underestimate the inverse scale parameter $f(A)$) when this variance is smaller than σ^2, the noise variance. This is undesirable, as we would like $f(A)$ to tend to infinity for irrelevant groups of variables. To circumvent this problem, we use an improper hyperprior of the form $p(f(A)) \propto f(A)^\beta$ to encourage $f(A)$ to go to infinity when the variance of $p(v_A)$ is smaller than σ^2. Consequently, the regularization term $-K\beta \sum_{A\in\mathcal{A}} \log f(A)$ with $\beta > 0$ is added to the objective function (11), and the only update that changes is that for $f(A)$:

$$f(A) = \frac{K(\beta + \frac{|A|}{2})}{\frac{1}{2}\sum_{k=1}^{K} \frac{1}{\zeta_A^k}(\|v_A^k\|_2^2 + \operatorname{Tr}\Sigma_{AA}^k)}. \tag{19}$$

Thus, we substitute the approximate type-II maximum likelihood estimation of $f(A)$ by approximate (also "type-II") maximum a posteriori estimation. In Sections 3.1 and 3.2 we empirically study the effect of the parameter β on the approximation quality.

3.1 Scale parameter inference with only one variable

In this experiment we evaluate the performance of the variational method described in Section 2.3 in recovering the unknown scale parameter f of the prior in the simplest, 1-dimensional case (note that in this subsection we omit the subscripts A as $\mathcal{A} = \{\{1\}\}$). More specifically, our goal here is to answer the following questions: Given an i.i.d. sample drawn from a univariate Student's t with shape and inverse scale parameters a and f, corrupted by Gaussian noise, and supposing we know both the noise variance σ^2 and the shape parameter a, can we precisely estimate the inverse scale parameter f using the variational method from Section 2.3? In the settings where we cannot, does regularization improve our estimates?

Experimental setup. We consider 10,000 tasks with one variable and one observation each (N, P, and X^k for all k equal to 1). Data are generated from the model with Student's t prior on v^k with parameters a set to 1.5 and f varying in the set $\mathcal{F}$ of 14 values between 0.02 and 50 taken roughly uniformly on the logarithmic scale, and Gaussian noise with variance σ^2 set to 1. We compare the performance of the variational method with that of a grid search over $\mathcal{F} \cup \{10^5\}$, where we use the trapezoidal rule to numerically solve the intractable integral in (6). The grid search, feasible in this basic setting, provides the best available approximation to the regularized maximum likelihood solution. To reduce the effect of random fluctuations, we repeat all experiments 5 times with different random seeds and report averaged results.

Results. Figure 3 summarizes the results. For three values of the parameter β, we plot (on the logarithmic scale) the estimated against the true variance for the considered range of the parameter f (recall that the variance of a Student's t-distributed random variable with parameters a and f equals $\frac{1}{(a-1)f}$). In all figures, we also plot the variance of the Gaussian noise σ^2. We observe that in the absence of regularization ($\beta = 0$) and when the signal is not much stronger than noise, the variational method overestimates the signal variance while the grid search does not. As we add regularization, this effect gradually goes away and the signal variance estimate is set to 0 (i.e., the estimate of f, $\widehat{f}$, goes to infinity) if the true signal variance is smaller than a certain threshold. When the regularization is too strong ($\beta = 0.25$), the estimated signal variance drops to 0 even when the signal is stronger than the noise, and the variance of the signal is heavily underestimated. With the right amount of regularization ($\beta = 0.05$ in this case) we observe the desired behavior: The variational method recovers the signal when it is stronger than noise, and sets $\widehat{f}$ to infinity otherwise. In all cases, variational estimates are close to the maximum likelihood estimates obtained by the grid search when the signal is much stronger than the noise.

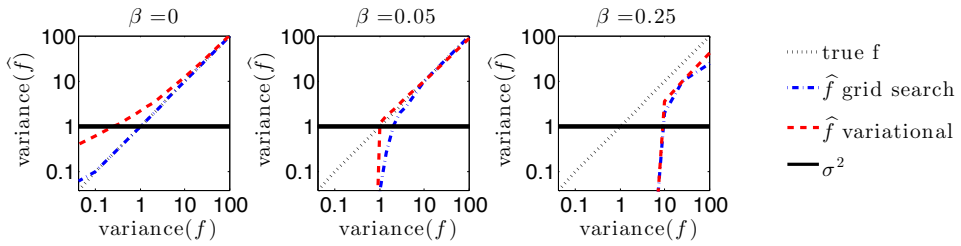

Figure 3: Recovery of the variance of the univariate Student's t distribution with added Gaussian noise of known variance with grid search and the variational method, with different levels of regularization. The x and y axes represent the variance based on the true and on the estimated f parameter values, respectively.

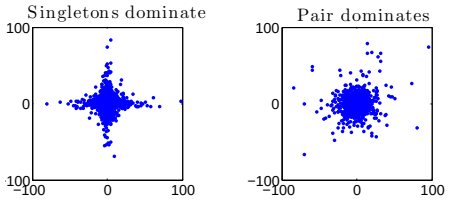

Figure 4: On the left, the singletons are the relevant groups. On the right, the pair is the relevant group.

3.2 Structured sparsity with two variables

In this section we empirically study the most basic case of the group relevance learning problem. Suppose that in each task we only have 2 variables, and therefore 3 possible groups, $\mathcal{A} = \{\{1\}, \{2\}, \{1,2\}\}$. Let X^k be the identity matrix in each task. In this basic setting, and supposing that the data come from the model, can our inference algorithm distinguish the case where the data $\{y^k\}_{k=1,\ldots,K}$ are generated by the group of variables $\{1,2\}$ from the opposite case, where the relevant groups are the two singletons $\{1\}$ and $\{2\}$?

These two settings differ in fact significantly in the case of a heavy-tailed prior on v_A^k: We have $w^k = v_{\{1\}}^k + v_{\{2\}}^k + v_{\{1,2\}}^k$; If $\{1,2\}$ is relevant and $\{1\}$ and $\{2\}$ are not, then $v_{\{1\}}^k$ and $v_{\{2\}}^k$ will have to be close to zero for all k, however, $v_{\{1,2\}}^k$ will be significantly far from zero for some k. As the prior on v_A^k only depends on v_A through its norm, these $v_{\{1,2\}}^k$ can be anywhere on the circle with radius $\|v_{\{1,2\}}^k\|_2$ with the same probability and therefore y^k can also be anywhere on the circle with radius $\|y^k\|_2$. In contrast, when $\{1,2\}$ is irrelevant and $\{1\}$ and $\{2\}$ are relevant, the rare events of $v_{\{1\}}$ and $v_{\{2\}}$ both being significantly away from zero will not occur at the same time for most k, and therefore the y^k with a large norm will tend to be concentrated along the axes. This behavior (using Student's t prior with parameter $a = 1.1$ on v_A^k) is illustrated in Figure 4, where we have plotted the data $\{y^k\}_{k=1,\ldots,K}$ for $K = 5,000$ in both settings.

Experimental setup. We consider 5,000 tasks with N and P equal to 2, with the set of groups $\mathcal{A} = \{\{1\}, \{2\}, \{1,2\}\}$. The data are generated from the model with Student's t prior on v^k with parameters a set to 1.5 and each $f(A)$ varying in a set of 14 values between 0.01 and 25 taken roughly uniformly on the logarithmic scale ($f(\{1\})$ and $f(\{2\})$ always equal each other), and Gaussian noise with variance σ^2 set to 1.

Results. Figure 5 summarizes the results for three values of the regularization parameter β ($\beta = 0$ corresponds to the absence of regularization). We report when the estimated pair variance $\frac{2}{(a-1)\widehat{f}(\{1,2\})}$ dominates (blue) or is dominated (red) by the estimated singleton variance

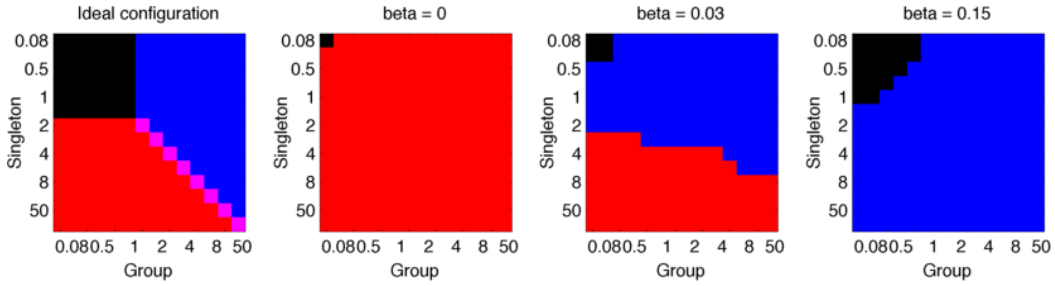

Figure 5: A red (blue) square means that the estimate of the singleton (group) variance is larger than the estimate of the group (singleton) variance for the corresponding true singleton and pair variances indicated by the axes. A black square means that both singleton and pair variances are under $2\sigma^2$, the noise variance. Best seen in color.

$\frac{1}{(a-1)\widehat{f}(\{1\})} + \frac{1}{(a-1)\widehat{f}(\{2\})}$, provided that one of them is larger than the noise variance, $2\sigma^2$. We see that when we do not regularize, the variational method explains everything with the singletons. As we add regularization, the pair explains more and more variance, however in such a way that the pair also explains the signal coming from singletons. Nonetheless, there is a regime ($\beta = 0.03$) where a strong signal coming from both the singletons and the pair is identified correctly. If we regularize too strongly ($\beta = 0.15$), the entire signal is explained by the pair, regardless of its source.

4 Experiments

In our experiments we consider two different instances of the denoising problem and we empirically evaluate the performance of our approach in recovering both the signal and the structure.

4.1 Structured sparsity in the context of denoising

In this section we study toy multi-task structured sparse denoising problems with more involved structure than those considered in the previous section. Our goal is to see whether our inference scheme is able to recover the structure, and whether learning the structure can help make better predictions. More specifically, we would like to answer the following questions: Given data $\{y^k\}_{k=1,\dots,K}$, generated from the model and supposing that we know the true shape parameter a of the Student's t and the noise variance σ^2, (a) can we recover the structure (i.e., the relevant groups and their weights), and (b) if we use the correct structure on previously unseen data from the same distribution, is our denoising more accurate than when using a different structure?

Experimental setup. To this end, we consider 1,000 tasks with N and P equal to 10, with the set of groups $\mathcal{A} = \{\{Q\}_{Q=1,\dots,P}, \{1,\dots,Q\}_{Q=2,\dots,P}\}$. Each signal w^k is generated using Student's t with parameters a set to 1.5 and $f(A)$ set to 0.2 or to 200, depending on whether A is considered relevant or irrelevant: In this fashion, the variance of the signal coming from relevant A is $\frac{|A|}{(a-1)f(A)} = 10 \times |A|$ (respectively, $0.01 \times |A|$ for irrelevant A). For each task k, y^k is a perturbed version of the signal w^k with additive Gaussian noise of variance $\sigma^2 I$.

We consider three different ways of generating data:

- **Singletons**: Here, only $\{1\},\dots,\{5\}$ are relevant, all other groups in $\mathcal{A}$ are irrelevant. The total signal variance therefore equals $5 \times 10 \times 1 = 50$.
- **One group**: Only $\{1,2,3,4,5\}$ is relevant. The variance equals 50 as well.
- **Overlapping groups**: The groups $\{1\}$, $\{1,2\}$, $\dots$, $\{1,2,3,4,5\}$ are relevant. The total signal variance equals $10 \times (1+2+3+4+5) = 150$.

	Singletons	One group	Overlapping
LASSO	18.53	18.04	56.26
W. LASSO	**14.66**	14.28	42.32
Structured	15.22	**13.6**	**42.04**

Table 1: Minimum average mean squared error for each combination of data generation and learning models.

For the three cases, we choose σ^2 to be 5, 5 and 15 respectively, so that the total noise variance $P\sigma^2$ equals the total signal variance in each case.

We also consider three nested models for inference:

- **LASSO-like**: In this simplest model, we only use the singletons, $\mathcal{A} = \{\{1\}, \dots, \{P\}\}$, and moreover, we force $f(A)$ to be constant across $\mathcal{A}$; In order to do so, we change the update for $f(A)$ to $f(A) = \frac{K \sum_{A \in \mathcal{A}} (\beta + \frac{|A|}{2})}{\frac{1}{2} \sum_{k=1}^{K} \sum_{A \in \mathcal{A}} \frac{1}{\zeta_A^k} (\|v_A^k\|_2^2 + \operatorname{Tr} \Sigma_{AA}^k)}$. This mimics the behavior of the LASSO, as the prior knowledge (that we are learning here) is the same for each coefficient.
- **Weighted LASSO-like**: The usual model with $\mathcal{A} = \{\{1\}, \dots, \{P\}\}$, with no further restrictions.
- **Structured**: The usual model with the full set of groups $\mathcal{A} = \{\{Q\}_{Q=1,\dots,P}, \{1, \dots, Q\}_{Q=2,\dots,P}\}$.

We examine each of the 9 combinations of data generation and learning models. In each case, we use 80% of the tasks for learning $\{f(A)\}_{A \in \mathcal{A}}$ and the rest for testing (we keep $\{f(A)\}_{A \in \mathcal{A}}$ fixed and infer the other parameters, v_A^k and ζ_A^k). We consider a range of 7 values for the regularization parameter β in each case. We repeat all experiments 5 times with different random seeds and report averaged results.

Results. We begin by examining the performance of each of the three models in *signal recovery*: In Table 1 we report the best average test error (i.e., the mean squared difference between the true and the learned signals w^k) for each of the 9 combinations. For all three regimes for data generation, the LASSO-like model performs far worse than the two others in recovery. This is due to the fact that this model learns the same prior for all variables, although each variable has a different marginal variance. In the first regime, when the data come from singletons, W.LASSO achieves better performance than the structured model. In the two other regimes, when the data come from one and multiple groups respectively, the Structured model outperforms W.LASSO.

In terms of *structure recovery*, for all three data generation regimes, we find one or more values of β that lead to the recovery of the relevant groups by Structured, with either the same or a slightly different β value leading to the smallest error in signal recovery. Figure 6 illustrates the percentage of total explained signal variance by each group for the Overlapping data generation regime and for the Structured model, for all considered regularization parameters: With no regularization, the model explains the signal with both the relevant groups and the singletons included in the relevant groups, however with more and more regularization, the signal variance explained by smaller groups is taken over by larger ones. The groups containing elements from $\{6, \dots, 10\}$, not shown in the plot, explain no variance in no regularization regime, with the exception of the largest group $\{1, \dots, 10\}$ that explains the weak signal coming from the irrelevant groups (recall that we have non-zero signal variance $0.01 \times |A|$ for the irrelevant groups A) in weak and moderate regularization regimes and takes over the whole signal variance when the regularization is too strong ($\beta \geq 0.15$).

In summary, in this set of experiments our inference algorithm is able to recover structure, and moreover, learning with the correct structure leads to better recovery.

4.2 Image denoising with wavelets

In this section we consider the image denoising problem using wavelets. The Haar wavelet basis for 2-dimensional images (Mallat, 1998) can naturally be arranged in a rooted directed tree where

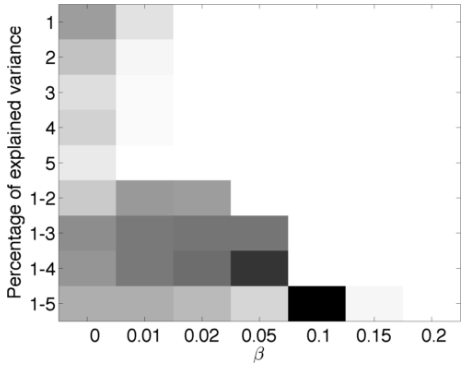

Figure 6: For each group of variables on the y axis, the intensity of gray indicates the percentage of total explained variance per β.

Image	Barbara	House
LASSO	138	72
W. LASSO	138	71
Structured	**134**	**58**

Table 2: Minimum average mean squared error for learning model on each image.

the structured sparsity-inducing norms with groups that are paths from the root have shown improvements over the ℓ_1 norm (Jenatton et al., 2011b).

Experimental setup. In order to denoise a large grayscale image, we cut it into small patches of 8×8 pixels, which compose the multiple 64-dimensional signals which we denoise simultaneously by learning the appropriate structured prior. Like in the previous section, we compare models with a uniform factorized sparse prior (LASSO), a non-uniform factorized sparse prior (W.LASSO), and the structured norms with all paths from the root of the tree to any node. We use two well-known images, Barbara and House (256×256 pixels each). Each signal w_k is formed by the wavelet coefficients of one 8×8 patch. For each of the $K = 1024$ tasks we form y^k by adding Gaussian noise of variance 400 along each dimension. As in the previous section, we use 80% of the tasks for learning $\{f(A)\}_{A \in \mathcal{A}}$ and the rest for testing.

Results. Table 2 shows the best test performance in terms of the mean squared error (divided by 100 for readability) of each method on each image. On both images, Structured outperforms both LASSO and W.LASSO. LASSO and W.LASSO yield comparable performance in recovery. Inspecting the relevances of different groups (paths in the wavelet tree) learned by Structured with the best regularization parameter, we see that the singletons other than the root wavelet explain less than 1% of the total explained variance; 99% are explained by the root and about 20 overlapping groups of 3 or 4 elements each.

5 Conclusions and Future Work

In this paper, we have proposed a flexible and general probabilistic model and an associated inference scheme for automatically learning the weights of possibly overlapping groups in the context of structured sparse multi-task linear regression. We have shown that the classical variational inference scheme is not well adapted for learning with this model, and have proposed a regularization method that closes this gap.

We have shown that our inference scheme is able to recover the model parameters with reasonable accuracy. In our future work, we plan to consider greedy active set approaches as by Bach (2008) with a view to making the inference in our model scalable to the real-world settings with large P and a large number of potential groups in $\mathcal{A}$.

We may also consider different likelihood models to handle settings different from linear regression, such as binary classification. Learning group relevances for classification is indeed crucial, e.g., in the context of genome-wide association studies with binary phenotypes in computational biology, or for image segmentation in computer vision.

In the appendix we provide details on the derivation of the variational inference scheme for our model (briefly introduced in Section 2.3) and discuss efficient ways of implementing the closed-form updates (12).

A Variational inference for the super-Gaussian structured sparse prior

In this appendix we derive step by step the variational updates given in Section 2.3.

We first recall our model: We assume that q_A is a *super-Gaussian* distribution, that is, the logarithm of $q_A(u)$ is convex in u^2 and non-increasing (Palmer et al., 2006). It therefore admits a representation of the following form by convex conjugacy

$$\log q_A(u) = \sup_{s \geq 0} -\frac{u^2}{2s} - \phi_A(s), \tag{7 revisited}$$

where $\phi_A(s)$ is convex in $1/s$. Note that the expression under the supremum in (7) has a unique maximizer that we denote $s^*(u)$. From (7), we get the following variational representation for $p(v_A^k|f(A))$:

$$\begin{aligned} p(v_A^k|f(A)) &= f(A)^{\frac{|A|}{2}} \sup_{\zeta_A^k \geq 0} e^{-\frac{||v_A^k||^2 f(A)}{2\zeta_A^k} - \phi_A(\zeta_A^k)} \\ &= f(A)^{\frac{|A|}{2}} \sup_{\zeta_A^k \geq 0} \Big[\mathcal{N}\left(v_A^k|0, \frac{\zeta_A^k}{f(A)} I\right) \left(2\pi \frac{\zeta_A^k}{f(A)}\right)^{\frac{|A|}{2}} e^{-\phi_A(\zeta_A^k)} \Big]. \end{aligned}$$

Finally, as v_A^k are assumed independent,

$$p(w^k|f) = \prod_{A \in \mathcal{A}} p(v_A^k|f(A)). \tag{20}$$

Our goal is to infer the set function f from data by maximizing the type II log-likelihood,

$$\sum_{k=1}^{K} \log p(y^k|f).$$

We tackle the problem using variational inference and consider the following lower bound, for sets $A \in \mathcal{A} \subseteq 2^V$ and for each regression task k, where we use the following notations:

- v^k is the concatenation of all elements indexed by elements of A in v_A^k, $A \in \mathcal{A}$,
- Z^k is a square diagonal matrix of dimension $\sum_{A \in \mathcal{A}} |A|$. Its diagonal consists of ζ_A^k, replicated $|A|$ times, for each $A \in \mathcal{A}$.
- F^k is a square diagonal matrix of dimension $\sum_{A \in \mathcal{A}} |A|$. Its diagonal consists of $f(A)$, replicated $|A|$ times, for each $A \in \mathcal{A}$.
- M is a matrix of dimension $P \times \sum_{A \in \mathcal{A}} |A|$ that ensures $w^k = M v^k$.

$$
\begin{aligned}
&\log p(y^k|f) \\
&\quad= \log \int_{\mathbb{R}^P} p(y^k|w^k)p(w^k|f)dw^k \\
&\quad= \log \int_{\mathbb{R}^P} \mathcal{N}(y^k|X^kMv^k,{\sigma^k}^2I) \prod_{A\in\mathcal{A}} \sup_{\zeta_A^k\geq 0} f(A)^{\frac{|A|}{2}} \Big[\mathcal{N}\Big(v_A^k|0,\frac{\zeta_A^k}{f(A)}I\Big)\Big(2\pi\frac{\zeta_A^k}{f(A)}\Big)^{\frac{|A|}{2}} e^{-\phi_A(\zeta_A^k)}\Big] \prod_{A\in\mathcal{A}} dv_A^k \\
&\quad= \log \int_{\mathbb{R}^P} \sup_{\zeta_A^k\geq 0, A\in\mathcal{A}} \mathcal{N}(y^k|X^kMv^k,{\sigma^k}^2I) \prod_{A\in\mathcal{A}} f(A)^{\frac{|A|}{2}} \mathcal{N}\Big(v_A^k|0,\frac{\zeta_A^k}{f(A)}I\Big)\Big(2\pi\frac{\zeta_A^k}{f(A)}\Big)^{\frac{|A|}{2}} e^{-\phi_A(\zeta_A^k)} \prod_{A\in\mathcal{A}} dv_A^k \\
&\quad\geq \sup_{\zeta_A^k\geq 0, A\in\mathcal{A}} \log \int_{\mathbb{R}^P} \mathcal{N}(y^k|X^kMv^k,{\sigma^k}^2I) \prod_{A\in\mathcal{A}} f(A)^{\frac{|A|}{2}} \mathcal{N}\Big(v_A^k|0,\frac{\zeta_A^k}{f(A)}I\Big)\Big(2\pi\frac{\zeta_A^k}{f(A)}\Big)^{\frac{|A|}{2}} e^{-\phi_A(\zeta_A^k)} \prod_{A\in\mathcal{A}} dv_A^k \\
&\quad= \sup_{\zeta_A^k\geq 0, A\in\mathcal{A}} \Big\{ \log \int_{\mathbb{R}^P} \mathcal{N}(y^k|X^kMv^k,{\sigma^k}^2I) \prod_{A\in\mathcal{A}} \mathcal{N}\Big(v_A^k|0,\frac{\zeta_A^k}{f(A)}I\Big) \prod_{A\in\mathcal{A}} dv_A^k \\
&\qquad\qquad + \sum_{A\in\mathcal{A}} \log\Big[f(A)^{\frac{|A|}{2}}\Big(2\pi\frac{\zeta_A^k}{f(A)}\Big)^{\frac{|A|}{2}} e^{-\phi_A(\zeta_A^k)}\Big]\Big\} \\
&\quad= \sup_{\zeta_A^k\geq 0, A\in\mathcal{A}} \Big\{ \log \int_{\mathbb{R}^P} \mathcal{N}(y^k|X^kMv^k,{\sigma^k}^2I)\mathcal{N}\Big(v^k|0,Z^kF^{-1}\Big)dv^k \\
&\qquad\qquad + \sum_{A\in\mathcal{A}} \Big[\frac{|A|}{2}\log f(A) + \frac{|A|}{2}\log\Big(2\pi\frac{\zeta_A^k}{f(A)}\Big) - \phi_A(\zeta_A^k)\Big]\Big\} \\
&\quad= \sup_{\zeta_A^k\geq 0, A\in\mathcal{A}} \Big\{ \log\mathcal{N}(y^k|0,X^kMZ^kF^{-1}M^T{X^k}^T + \sigma^2I) \\
&\qquad\qquad + \sum_{A\in\mathcal{A}} \Big[\frac{|A|}{2}\log f(A) + \frac{|A|}{2}\log\Big(2\pi\frac{\zeta_A^k}{f(A)}\Big) - \phi_A(\zeta_A^k)\Big]\Big\} \\
&\quad= \sup_{\zeta_A^k\geq 0, A\in\mathcal{A}} \Big\{ -\frac{1}{2}{y^k}^T\Big(X^kMZ^kF^{-1}M^T{X^k}^T + \sigma^2I\Big)^{-1}y^k - \frac{1}{2}\log\det\Big(X^kMZ^kF^{-1}M^T{X^k}^T + \sigma^2I\Big) \\
&\qquad\qquad - \frac{N}{2}\log(2\pi) + \sum_{A\in\mathcal{A}} \Big[\frac{|A|}{2}\log f(A) + \frac{|A|}{2}\log\Big(2\pi\frac{\zeta_A^k}{f(A)}\Big) - \phi_A(\zeta_A^k)\Big]\Big\}.
\end{aligned}
$$

Thus, we need to minimize the following overall bound with respect to f and ζ_A^k for all $A\in\mathcal{A}$ and $k\in\{1,\ldots,K\}$:

$$
\begin{aligned}
-\sum_{k=1}^K \Big\{ &-\frac{1}{2}{y^k}^T\Big(X^kMZ^kF^{-1}M^T{X^k}^T + \sigma^2I\Big)^{-1}y^k - \frac{1}{2}\log\det\Big(X^kMZ^kF^{-1}M^T{X^k}^T + \sigma^2I\Big) \\
&+ \sum_{A\in\mathcal{A}} \frac{|A|}{2}\log f(A) + \frac{\sum_{A\in\mathcal{A}}|A| - N}{2}\log(2\pi) + \frac{1}{2}\log\det(Z^kF^{-1}) - \sum_{A\in\mathcal{A}}\phi_A(\zeta_A^k)\Big\}.
\end{aligned}
\tag{10 revisited}
$$

In its form given by (10), the bound is difficult to optimize. However, we can recognize parts of it as minima of convex functions, which will allow us to design an iterative algorithm with analytic updates, finding a local minimum. In particular, it is not difficult to show that

$$
\begin{aligned}
&\frac{1}{2}\log\det\Big(X^kMZ^kF^{-1}M^T{X^k}^T + \sigma^2I\Big) - \frac{1}{2}\log\det(Z^kF^{-1}) \\
&\qquad= \frac{1}{2}\log\det\Big(M^T{X^k}^TX^kM + \sigma^2FZ^{k^{-1}}\Big) + \frac{N - \sum_{A\in\mathcal{A}}|A|}{2}\log(\sigma^2) \\
&\qquad= \inf_{\Lambda^k\succcurlyeq 0} \frac{1}{2}\operatorname{Tr}\{(M^T{X^k}^TX^kM + \sigma^2FZ^{k^{-1}})\Lambda^k\} - \frac{1}{2}\log\det\Lambda^k - \frac{1}{2}\sum_{A\in\mathcal{A}}|A| + \frac{N - \sum_{A\in\mathcal{A}}|A|}{2}\log(\sigma^2),
\end{aligned}
$$

which, with a change of variables $\Sigma^k = \sigma^2 \Lambda^k$, is written as

$$\inf_{\Sigma^k \succcurlyeq 0} \frac{1}{2\sigma^2} \operatorname{Tr}\{(M^T {X^k}^T X^k M)\} + \frac{1}{2} \sum_{A \in \mathcal{A}} \frac{\operatorname{Tr} \Sigma^k f(A)}{\zeta_A^k} - \frac{1}{2} \log\det \Sigma^k - \frac{1}{2} \sum_{A \in \mathcal{A}} |A| + \frac{N}{2} \log(\sigma^2), \quad (21)$$

and that

$$\frac{1}{2} {y^k}^T \Big(X^k M Z^k F^{-1} M^T {X^k}^T + \sigma^2 I \Big)^{-1} y^k = \inf_{v^k} \frac{1}{2\sigma^2} ||y^k - X^k M v^k||_2^2 + \frac{{v^k}^T F {Z^k}^{-1} v^k}{2}.$$

Thus, from (10), our optimization problem becomes

$$\begin{aligned} \inf_{\zeta^k \geq 0} \inf_{v^k} \inf_{\Sigma^k \succcurlyeq 0} \sum_{k=1}^K \Big\{ & \frac{1}{2\sigma^2} ||y^k - X^k M v^k||_2^2 + \frac{1}{2} \sum_{A \in \mathcal{A}} \frac{f(A)}{\zeta_A^k} (||v_A^k||_2^2 + \operatorname{Tr} \Sigma_{AA}^k) \\ & - \frac{1}{2} \log\det \Sigma^k + \frac{N}{2} \log(\sigma^2) + \frac{N}{2} \log(2\pi) + \frac{1}{2\sigma^2} \operatorname{Tr} M^T {X^k}^T X^k M \Sigma^k - \frac{1}{2} \sum_{A \in \mathcal{A}} |A| \\ & + \sum_{A \in \mathcal{A}} \Big[- \frac{|A|}{2} \log 2\pi + \phi_A(\zeta_A^k) - \frac{1}{2} |A| \log f(A) \Big] \Big\}, \end{aligned} \quad \text{(11 revisited)}$$

and the updates are

$$\begin{aligned} \Sigma^k &= \sigma^2 (M^T {X^k}^T X^k M + \sigma^2 F {Z^k}^{-1})^{-1} \\ v^k &= (M^T {X^k}^T X^k M + \sigma^2 F {Z^k}^{-1})^{-1} M^T {X^k}^T y^k \\ \zeta_A^k &= \operatorname*{argmin}_{z \geq 0} \phi_A(z) + \frac{1}{2} \frac{f(A)}{z} (||v_A^k||_2^2 + \operatorname{Tr} \Sigma_{AA}^k) \\ \sigma^2 &= \frac{\sum_{k=1}^K \{ ||y^k - X^k M v^k||_2^2 + \operatorname{Tr} M^T {X^k}^T X^k M \Sigma^k \}}{NK} \\ f(A) &= \operatorname*{argmin}_x \frac{x}{2} \sum_{k=1}^K \frac{1}{\zeta_A^k} (||v_A^k||_2^2 + \operatorname{Tr} \Sigma_{AA}^k) - K \frac{1}{2} |A| \log x \\ &= \frac{K|A|}{\sum_{k=1}^K \frac{1}{\zeta_A^k} (||v_A^k||_2^2 + \operatorname{Tr} \Sigma_{AA}^k)}. \end{aligned} \quad (22)$$

A.1 Regularized version

As we empirically show in Section 3.1 of the main paper, the variational approximation scheme from above tends to overestimate the variance of the prior distribution (i.e., underestimate the inverse scale parameter $f(A)$) when this variance is smaller than σ^2, the noise variance. This is undesirable, as we would like $f(A)$ to go to infinity for irrelevant subsets of variables. To circumvent this problem, we use an improper prior of the form

$$p(f(A)) \propto f(A)^\beta$$

to encourage $f(A)$ to go to infinity when the variance of $p(v_A)$ is smaller than σ^2. Consequently, the term $-\beta \log f(A)$ is added to the objetive function (11), and the only update that changes is the update for $f(A)$:

$$\begin{aligned} f(A) &= \operatorname*{argmin}_x \frac{x}{2} \sum_{k=1}^K \frac{1}{\zeta_A^k} (||v_A^k||_2^2 + \operatorname{Tr} \Sigma_{AA}^k) - K(\frac{|A|}{2} + \beta) \log x \\ &= \frac{K(\frac{|A|}{2} + \beta)}{\frac{1}{2} \sum_{k=1}^K \frac{1}{\zeta_A^k} (||v_A^k||_2^2 + \operatorname{Tr} \Sigma_{AA}^k)}. \end{aligned} \quad (23)$$

A.2 Faster updates

The update equations (22) involve the inversion of the $\sum_{A\in\mathcal{A}}|A|\times\sum_{A\in\mathcal{A}}|A|$ matrix $M^T{X^k}^TX^kM+\sigma^2FZ^{k^{-1}}$. In fact, using the matrix inversion lemma, we can avoid performing this expensive inversion by rewriting the updates so that we only have to invert a $P\times P$ or an $N\times N$ matrix instead.

Before we write down the modified updates, let us introduce some additional shorthand notation:

- $\xi^k\in\mathbb{R}^P$ is the sum $\sum_{A\in\mathcal{A}}\frac{\zeta_A^k}{f(A)}1_A$, where $1_A\in\mathbb{R}^P$ denotes the indicator vector for the index set A;
- $\Xi^k\in\mathbb{R}^{P\times P}$ is a square diagonal matrix with $\Xi^k_{ii}=\xi^k_i, i=1,\dots,P$; Put differently, $\Xi^k=MZ^kF^{-1}M^\top$.
- H is a square diagonal matrix corresponding to Z^kF^{-1}.

$P\times P$ matrix inversion.

$$\begin{aligned}
\Sigma^k &= H - HM^\top\Xi^{-1}MH + \sigma^2HM^\top\Xi^{-1}\big({X^k}^\top X^k+\sigma^2\Xi^{-1}\big)^{-1}\Xi^{-1}MH\\
\operatorname{Tr}\Sigma^k_{AA} &= |A|\frac{\zeta_A^k}{f(A)} - \frac{{\zeta_A^k}^2}{f(A)^2}\sum_{i\in A}\frac{1}{\xi_i} + \sigma^2\frac{{\zeta_A^k}^2}{f(A)^2}\sum_{i,j\in A}\frac{1}{\xi_i\xi_j}\big[\big({X^k}^\top X^k+\sigma^2\Xi^{-1}\big)^{-1}\big]_{ij}\\
v^k &= HM^\top\Xi^{-1}\big({X^k}^\top X^k+\sigma^2\Xi^{-1}\big)^{-1}{X^k}^\top y^k\\
\|v^k_A\|_2^2 &= \frac{{\zeta_A^k}^2}{f(A)^2}\sum_{i\in A}\frac{1}{\xi_i^2}\big[\big({X^k}^\top X^k+\sigma^2\Xi^{-1}\big)^{-1}{X^k}^\top y^k\big]_i^2.
\end{aligned}\tag{24}$$

$N\times N$ matrix inversion.

$$\begin{aligned}
\Sigma^k &= H - HM^\top{X^k}^\top\big(X^k\Xi^k{X^k}^\top+\sigma^2I\big)^{-1}X^kMH\\
v^k &= HM^T{X^k}^T\big(X^k\Xi^k{X^k}^\top+\sigma^2I\big)^{-1}y^k.
\end{aligned}\tag{25}$$

Special case of no design (signal denoising). When $X^k=I$, the computations become considerably simpler. Note that in this case $N=P$ and the matrix ${X^k}^\top X^k+\sigma^2\Xi^{-1}$ is diagonal, so the cost of its inversion is $O(P)$ instead of $O(P^3)$. In fact, we do not even need to form the diagonal matrix if we do not explicitly use Σ^k and v^k. The updates can be rewritten as follows:

$$\begin{aligned}
\operatorname{Tr}\Sigma^k_{AA} &= |A|\frac{\zeta_A^k}{f(A)} - \frac{{\zeta_A^k}^2}{f(A)^2}\sum_{i\in A}\frac{1}{\xi_i} + \sigma^2\frac{{\zeta_A^k}^2}{f(A)^2}\sum_{i\in A}\frac{1}{\xi_i^2}\frac{1}{(1+\sigma^2/\xi_i)}\\
\|v^k_A\|_2^2 &= \frac{{\zeta_A^k}^2}{f(A)^2}\sum_{i\in A}\frac{1}{\xi_i^2}\Big[\frac{y_i^k}{1+\sigma^2/\xi_i}\Big]^2\\
\zeta_A^k &= \operatorname*{argmin}_{z\ge0}\,\phi_A(z)+\frac{1}{2}\frac{f(A)}{z}(||v^k_A||_2^2+\operatorname{Tr}\Sigma^k_{AA})\\
f(A) &= \frac{K(\frac{|A|}{2}+\beta)}{\frac{1}{2}\sum_{k=1}^K\frac{1}{\zeta_A^k}(||v^k_A||_2^2+\operatorname{Tr}\Sigma^k_{AA})}.
\end{aligned}\tag{26}$$

(The updates for $f(A)$ and ζ_A^k remain unchanged.)

Acknowledgements

This work was supported by the European Research Council (SIERRA project 239993). The authors would like to thank Guillaume Obozinski for fruitful discussions, Julien Mairal for his advice on setting up experiments on images, and Sylvain Arlot for his comments on the manuscript.

References

Azencott, C.-A., Grimm, D., Sugiyama, M., Kawahara, Y., and Borgwardt, K. M. Efficient network-guided multi-locus association mapping with graph cut. *Bioinformatics*, 29(13):i171–i179, 2013.

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2012a.

Bach, F. Exploring large feature spaces with hierarchical multiple kernel learning. In *Advances in Neural Information Processing Systems*, 2008.

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. Structured sparsity through convex optimization. *Statistical Science*, 27(4):450–468, 2012b.

Bühlmann, P. and van de Geer, S. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer, 2011.

Cevher, V., Duarte, M. F., Hegde, C., and Baraniuk, R. G. Sparse signal recovery using Markov random fields. In *Advances in Neural Information Processing Systems*, volume 20, 2008.

Hernández-Lobato, D. and Hernández-Lobato, J. M. Learning feature selection dependencies in multi-task learning. In *Advances in Neural Information Processing Systems*, 2013.

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

Jacob, L., Obozinski, G., and Vert, J.-P. Group Lasso with overlap and graph Lasso. In *Proceedings of the International Conference on Machine Learning*, 2009.

Jenatton, R., Audibert, J.-Y., and Bach, F. Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research*, 12:2777–2824, 2011a.

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

Kim, S. and Xing, E. P. Tree-guided group lasso for multi-task regression with structured sparsity. In *Proceedings of the International Conference on Machine Learning*, 2010.

Mallat, S. *A Wavelet Tour of Signal Processing*. Academic Press, 1998.

Obozinski, G. and Bach, F. Convex relaxation for combinatorial penalties. Technical Report hal-00694765, May 2012.

Palmer, J. A., Wipf, D. P., Kreutz-Delgado, K., and Rao, B. D. Variational EM algorithms for non-gaussian latent variable models. In *Advances in Neural Information Processing Systems*, 2006.

Pascal, F., Bombrun, L., Tourneret, J.-Y., and Berthoumieu, Y. Parameter estimation for multivariate generalized gaussian distributions. *IEEE Transactions on Signal Processing*, 61(23): 5960–5971, 2013.

Rapaport, F., Zinovyev, A., Dutreix, M., Barillot, E., and Vert, J.-P. Classification of microarray data using gene networks. *BMC Bioinformatics*, 8, 2007.

Seeger, M. and Nickisch, H. Large scale bayesian inference and experimental design for sparse linear models. *SIAM Journal on Imaging Sciences*, 4(1):166–199, 2011.

Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006.

Zhao, P., Rocha, G., and Yu, B. The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist*, 37:3468–3497, 2009.