Incorporating Grouping Information into Bayesian Decision Tree Ensembles

Junliang Du [1] **Antonio R. Linero** [1]

Abstract

We consider the problem of nonparametric regression in the high-dimensional setting in which $P \gg N$. We study the use of overlapping group structures to improve prediction and variable selection. These structures arise commonly when analyzing DNA microarray data, where genes can naturally be grouped according to genetic pathways. We incorporate overlapping group structure into a Bayesian additive regression trees model using a prior constructed so that, if a variable from some group is used to construct a split, this increases the probability that subsequent splits will use predictors from the same group. We refer to our model as an overlapping group Bayesian additive regression trees (OG-BART) model, and our prior on the splits an overlapping group Dirichlet (OG-Dirichlet) prior. Like the sparse group lasso, our prior encourages sparsity both within and between groups. We study the correlation structure of the prior, illustrate the proposed methodology on simulated data, and apply the methodology to gene expression data to learn which genetic pathways are predictive of breast cancer tumor metastasis.

1. Introduction

Modern datasets often have numbers of predictors which greatly exceed the number of observations. This complicates traditional tasks, such as identifying important predictors or forming predictions. To make headway on such problems, it is essential that the data exhibit some additional structure. Among such structures is the *sparsity* structure, in which the response depends on a small number of predictors (Zou & Hastie, 2005; Candes & Tao, 2007). When external data sources are available one can leverage this additional information to help learn these structures. For example, when covariates are associated to known graphical structures, such as gene regulatory networks, one can use this information to aid variable selection (Tibshirani & Taylor, 2011; Li & Zhang, 2010; Chang et al., 2016).

In this paper, we focus on incorporation of *grouping* information to learn sparse structures. Many tools for utilizing grouping structures have been developed for linear models; the canonical example is the group lasso (Yuan & Lin, 2006), which builds a penalty using the grouping structure as $\lambda \sum_{g=1}^{G} \sqrt{P_g} \|\beta_g\|_2$ where β_g denotes a collection of regression coefficients belonging to group g, $\|x\|_2 = (\sum_i x_i^2)^{1/2}$ is the ℓ_2 norm, and P_g is the number of predictors in group g. This approach is easily generalized, leading to, e.g., grouped SCAD estimators (Wang et al., 2007). Selection can be done at two separate levels: selection of groups and selection of predictors-within-groups. The group lasso gives selection at the group level, while the sparse group lasso gives selection at both levels simultaneously (Simon et al., 2013). Groups in general can be *overlapping* (with a predictor belonging to potentially more than one group) or *non-overlapping* (with a predictor belonging to at-most one group). Overlapping group structures have also been extensively studied in the context of linear models, leading to tools such as the overlapping group lasso (or SCAD) (Jacob et al., 2009). Bayesian approaches, which incorporate grouping information into an informative prior to select both groups and predictors-within-groups, also abound (Rockova & Lesaffre, 2014; Xu & Ghosh, 2015; Stingo & Vannucci, 2011).

There has been less attention paid to nonparametric regression models in the literature. In the context of genomic data, flexible models are of interest due to the fact that genetic effects are potentially nonlinear and may possess complex interaction structures (Basu et al., 2018). We take a Bayesian nonparametric approach to incorporating grouping information by embedding the grouping information into a suitably-specified prior over the trees in a Bayesian additive regression trees (BART) ensemble. The BART framework, which gives a Bayesian variant of the popular random forests (Breiman, 2001) and boosting (Freund et al., 1999) algorithms, has become increasingly popular in recent years due to its ease-of-use and high performance as a general purpose prediction tool. BART has become particularly popular in causal inference settings, where it consistently

[*]Equal contribution [1]Department of Statistics, Florida State University, Tallahassee, Florida. Correspondence to: Antonio R. Linero <arlinero@stat.fsu.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

performs well for both prediction and uncertainty quantification in the Atlantic Causal Inference Conference's annual data challenge (Dorie et al., 2017). Variants of BART have recently been developed for survival analysis (Sparapani et al., 2016), loglinear models (Murray, 2017), functional data analysis (Starling et al., 2018), and interaction detection (Du & Linero, 2019), among many other applications.

Previous analyses of the BART prior have suggested that a promising strategy for performing variable selection is to place a sparsity-inducing prior on the splitting probability vector $s = (s_1, \ldots, s_P)$, such that predictor j is selected to construct a split a-priori with probability s_j. This encodes sparsity into the model because, if s is (nearly) sparse, then relatively few predictors in the model will be used to build splits in the ensemble. Linero (2018) used a sparsity-inducing Dirichlet prior on s to construct a prior over the tree ensemble which splits on a small number of predictors, and showed that this approach can be used to perform automatic relevance determination (Neal, 1995) in high dimensional settings. The theoretical properties of this approach were studied by Linero & Yang (2018), who established posterior concentration of the resulting BART posterior distribution at within a logarithmic factor of the oracle minimax rate in the high dimensional setting with $\log P$ growing nearly as fast as N, even when the smoothness index α of the regression function and number of relevant predictors D are both unknown (similar results are also obtained by Rockova & van der Pas, 2017). Moreover, they show that BART has the attractive theoretical property of automatically adapting to low order interactions in the data.

Leveraging this strategy, we propose an overlapping group BART model (OG-BART) which naturally incorporates the grouping information into the model using what we refer to as an *overlapping group (OG) Dirichlet* prior. The OG-Dirichlet prior handles the overlapping and non-overlapping structures in a convenient, unified, setting. This prior has several close relatives in the literature, although we are not aware of any models with precisely the same structure. In the special case of non-overlapping groups, this reduces to a Dirichlet-tree distribution described by Minka (1999) and Dennis (1991), which is conjugate to multinomial sampling. In the overlapping setting, the OG-Dirichlet distribution is similar in structure to latent Dirichlet allocation (LDA) models (Blei et al., 2003) in which the individual groups correspond to "topics" and predictors as "words", but differs in that much of the within-topic sparsity is determined by the grouping information.

A use case of interest for grouped variable selection is to help select relevant genes, or pathways of genes, when analyzing gene expression data. In this case, one can build overlapping groups from the genetic pathways obtained from, for example, the KEGG database (Kanehisa & Goto, 2000). We illustrate our methodology first on simulated data in which the response depends non-linearly on the predictors, and demonstrate consistent gains when a correctly-specified grouping structure is known, while simultaneously having no loss in performance when an incorrect grouping structure is applied. We then apply our methodology to the breast cancer dataset compiled by Van De Vijver et al. (2002) to identify genes and pathways of genes whose expression levels are predictive of breast cancer tumor metastasis.

In Section 2, we describe the Bayesian additive regression trees framework and review how to obtain sparsity by placing a prior on the splitting proportions of the ensemble. In Section 3, we introduce our overlapping group Dirichlet prior, study the correlation structure in the asymptotic regime where the number of groups and predictors diverges, and discuss specification of hyperparameters. In Section 4, we illustrate the benefits of incorporating the grouping information in both the overlapping and non-overlapping settings, particularly focusing on the gains as the signal-to-noise ratio decreases. In Section 5 we apply our methodology to analyze the breast cancer dataset of Van De Vijver et al. (2002). We conclude in Section 6 with a discussion.

2. Review of BART Under Sparsity

2.1. Model and Notation

Let $\mathcal{D} = \{X_i, Y_i\}_{i=1}^N$ consist of N iid replicates of a predictor vector $X_i \in [0,1]^P$ and a response $Y_i \in \mathbb{R}$. For the moment, we focus on the nonparametric regression model $Y_i = f(X_i) + \epsilon_i$ where $\epsilon_i \overset{\text{iid}}{\sim} \text{Normal}(0, \sigma^2)$. Additionally, we assume that the coordinates of $X_i = (X_{i1}, \ldots, X_{iP})$ are associated to some number of G of *groups*. The grouping structure can be represented by a sparse binary matrix $M \in \mathbb{R}^{G \times P}$ such that $M_{gj} = 1$ if predictor j lies in group g and $M_{gj} = 0$ otherwise. We assume that the each predictor is associated to at least one group, i.e., $\sum_g M_{gj} \geq 1$, with the non-overlapping setting corresponding to $\sum_g M_{gj} = 1$ for all j. We assume that M is known a-priori.

A decision tree consists of a binary tree $\mathcal{T}$, with leaf nodes $\mathcal{L}$ and branch nodes $\mathcal{B}$, such that each branch node $b \in \mathcal{B}$ is associated to a decision rule of the form $[x_{j(b)} \leq C_b]$, as well as a left and right child node. Each leaf node $\ell \in \mathcal{L}$ is associated to a prediction μ_ℓ. For each branch b, if x is associated to b then it is further associated to the left or right child according as x satisfies the decision rule or not. We denote the collection of all leaf node parameters as $\mathcal{M}$. The tree $\mathcal{T}$ then naturally corresponds to a partition of $[0,1]^P$ and to a stepwise-constant function $g(\cdot; \mathcal{T}, \mathcal{M})$ such that $g(x; \mathcal{T}, \mathcal{M}) = \mu_\ell$ whenever x is associated to ℓ.

The BART framework, introduced in the seminal work of Chipman et al. (2010), models f as a sum of regression trees. We write $f \sim \text{BART}$ to denote a BART prior on

f, which is described by the following generative scheme. Draws from the prior $f \sim \text{BART}$ can be represented as $f(x) = \sum_{t=1}^{T} g(x; \mathcal{T}_t, \mathcal{M}_t)$ where $\mathcal{T}_t$ is a *random decision tree* and $\mathcal{M}_t$ denotes the parameters associated to the leaf nodes of $\mathcal{T}_t$. According to this prior, the pairs $(\mathcal{T}_t, \mathcal{M}_t)$ are generated independently (conditional on hyperparameters) as follows:

(i) Draw $\mathcal{T}_t$ from some distribution $\pi_{\mathcal{T}}$. This is usually a branching process in which, sequentially at each depth d, each node of depth d becomes a leaf node with probability $\gamma(1+d)^{-\beta}$ and becomes a branch node otherwise, but other options exist (Lakshminarayanan et al., 2014; Denison et al., 1998).

(ii) Conditional on $\mathcal{T}_t$, draw the leaf parameters $\mathcal{M}_t$ independently as $\mu_{t\ell} \overset{\text{iid}}{\sim} \pi_\mu$. For computational purposes, π_μ is usually a $\text{Normal}(0, \sigma_\mu^2/T)$ distribution, but can vary depending on the structure of the response (Murray, 2017; Pratola et al., 2017; Linero et al., 2018).

We emphasize that this describes draws from the prior; sampling from the posterior is typically carried out via Markov chain Monte Carlo (Chipman et al., 2010). Additionally, in view of Linero & Yang (2018), throughout this work we will replace the decision trees in the ensemble with "soft" decision trees (Irsoy et al., 2012) to improve predictive accuracy when the underlying function is smooth. We assume the splitting rule $[x_{j(b)} \leq C_b]$ is sampled from the prior as follows. First, the coordinate $j(b)$ used in the splitting rule is drawn according to a probability vector $s = (s_1, \dots, s_P)$. Second, we set $C_b \sim \text{Uniform}(L_{j(b)}, U_{j(b)})$ where (L_j, U_j) define the minimal/maximal values of x_j which can lead to branch b.

More detailed descriptions of BART can be found in Kapelner & Bleich (2016), Chipman et al. (2010), Chipman et al. (2013), or Linero (2017). A benefit of the BART framework is that there exist hyperparameter values which generalize surprisingly well across problems. Unless otherwise stated, we will use the default hyperparameter and hyperprior settings described by Linero & Yang (2018).

2.2. Dirichlet Priors on Splitting Proportions

Our starting point for incorporating grouping structure into the model is the observation that sparsity can be encoded in $s = (s_1, \dots, s_P)$ (the prior probabilities of splitting on each predictor). Without encoding any preference for sparsity in the model — say, by fixing $s_j = P^{-1}$ — Linero (2018) showed that the BART prior encourages many predictors to be relevant, but to have only a small influence on the response. In a large P asymptotic regime, this concentrates the prior on ensembles in which there are exactly as many predictors as branches. We refer to this as the *many weak effects* scenario, which is at odds with our desired assumption of sparsity. To see how s can be used to encode sparsity, if $P = 4$ and $s = (1/2, 1/2, 0, 0)$, then the resulting $f \sim \text{BART}$ will depend on, at most, two predictors. Exact sparsity, however, is not necessary, and we consider the conditionally conjugate sparsity-inducing Dirichlet prior $s \sim \text{Dirichlet}(\alpha/P, \dots, \alpha/P)$. While s will not be *exactly* sparse in this setting, this Dirichlet prior concentrates tightly in neighborhoods of sparse vectors (Yang & Dunson, 2014), which is sufficient to encourage exact sparsity in $f(\cdot)$.

Because of the conjugacy of the Dirichlet prior to multinomial sampling, this prior can be incorporated easily into existing Gibbs samplers by sampling from the full-conditional of s, which (under our specific choice of $\pi_{\mathcal{T}}$) is given by $s \sim \text{Dirichlet}(\alpha/P + c_1, \dots, \alpha/P + c_P)$ where c_j denotes the number of branch nodes in the ensemble which split on coordinate j. Additionally, one can tune the prior to attain a desired level of sparsity. It can be shown that, conditional on the number of splitting rules B in the ensemble, the prior on the number of predictors in the ensemble D is given approximately by $D - 1 \sim \text{Poisson}(\alpha \sum_{i=0}^{B-1} (\alpha + i)^{-1})$ (Linero, 2018).

Unfortunately, the Dirichlet prior is not suitable when we desire *positive* correlations in the s_j's. As noted, for example, by Blei & Lafferty (2006), the Dirichlet prior can encode only negative correlations among the s_j's. This conflicts with our desire to take advantage of grouping structure: if predictor j is included in the model, and lies in the same group as predictor k, we would like to increase rather than decrease the probability of k being included in the model.

2.3. BART for Classification

The BART framework can be extended to classification by using the data augmentation strategy of Albert & Chib (1993). If the response Y_i is binary, we introduce a latent variable R_i such that $Y_i = I(R_i > 0)$. The R_i's are then modeled as $R_i = f(X_i) + \epsilon_i$ where $\epsilon_i \sim \text{Normal}(0, 1)$. This induces a probit model, where $[Y_i \mid X_i] \sim \text{Bernoulli}(\pi_i)$ where $\pi_i = \Phi(f(X_i))$ and $\Phi(\cdot)$ is the distribution function of a standard Gaussian random variable. This model can be fit by adding the data augmentation step $R_i \sim \text{Normal}(f(X_i), 1)$ (truncated according to the value of Y_i), and subsequently using R_i as the response in all other steps of the Markov chain Monte Carlo algorithm described in the supplementary material.

3. Overlapping Group Priors

The overlapping group Dirichlet (OG-Dirichlet) prior is a generalization of the Dirichlet prior which allows for positive correlations among the s_j's when they lie in the same group. Recall that $M \in \mathbb{R}^{G \times P}$ denotes a

sparse binary matrix such that $M_{gj} = 1$ if predictor j lies in group g. The OG-Dirichlet prior sets $s_j = \sum_{g=1}^{G} \pi_g \omega_{gj}$ where π_g and ω_{gj} are assigned independent Dirichlet priors $\omega_g \sim \text{Dirichlet}(\psi q_{g1}, \ldots, \psi q_{gP})$ and $\pi \sim \text{Dirichlet}(\alpha\lambda_1, \ldots, \alpha\lambda_G)$. Equivalently, we can write $s = W\pi$ where column g of W is given by $(\omega_{g1}, \ldots, \omega_{gP})$. This corresponds to the following generative procedure for sampling splitting rules: first, sample a group g according to π; second, sample a variable j within group g according to $(\omega_{g1}, \ldots, \omega_{gP})$.

We take the coefficients $\{q_{gj}\} = Q \in \mathbb{R}^{G\times P}$ and $\lambda = (\lambda_1, \ldots, \lambda_G)$ to be fixed a-priori such that $\sum_g \lambda_g = \sum_j q_{gj} = 1$. Additionally, we take the coefficient q_{gj} to be non-zero only when $M_{gj} = 1$, with the understanding that (say) the third component of a $\text{Dirichlet}(1/2, 1/2, 0)$ random vector has a point-mass distribution at 0. Hence W will have the same sparsity pattern as M.

The OG-Dirichlet prior allows for sparsity at two separate levels. First, the model will be sparse at the between-group level when the parameter α is small. Second, the model will be sparse at the within-group level when the parameter ψ is small. As we will see in the simulation studies of Section 4, the biggest benefits of our model occur when there is high between-group sparsity and low within-group sparsity, although we find benefits regardless of the level of within-group sparsity.

The OG-Dirichlet prior is similar in structure to the latent Dirichlet allocation (LDA) model (Blei et al., 2003), with π playing the role of a distribution over topics and the columns of W playing the role of the distribution of words within each topic. Unlike LDA, we have further information on the structural sparsity of the matrix W. This connection with LDA, combined with the generative scheme described for sampling j described above, gives one strategy for updating (W, π) in a Gibbs sampler: we introduce latent variables Z_b for each branch b such that $Z_b = g$ if b was selected to be drawn according to group g. Conditional on the Z_b's, we now can take advantage of the conjugacy of the Dirichlet distribution to update W and π. A description of one possible Gibbs sampler is given in the supplementary material.

3.1. Correlation Structure of the Prior

The primary problem with the Dirichlet prior for our purposes is that it does not allow for positive correlation among components of s. Intuitively, we expect that the OG-Dirichlet prior should bypass this by inducing correlations between s_j and s_k if predictors j and k share groups. Roughly speaking, this correlation is determined by the magnitude of ψ and the proportion of groups which in which predictors j and k overlap. To make this precise, it is useful to consider an asymptotic regime in which the number of groups G and the number of predictors P are both large, which is highly typical in practice. As $(G, P) \to \infty$ we let $\lambda_g = \widetilde{\lambda}_g/G$ with λ_g fixed, but impose that each predictor j lies in finitely many groups with q_{gj} fixed. Under these conditions, the following result holds for the correlation structure (a proof is given in the supplementary material).

Proposition 3.1. *Under the asymptotic regime described in the previous paragraph, we have*

$$\text{Cor}(s_j, s_k) \sim \frac{\psi q_j^\top \Lambda q_k}{\sqrt{\psi q_j^\top \Lambda q_j + q_j^\top \Lambda \mathbf{1}} \cdot \sqrt{\psi q_k^\top \Lambda q_k + q_k^\top \Lambda \mathbf{1}}},$$

where $\Lambda = \text{diag}(\widetilde{\lambda}_1, \widetilde{\lambda}_2, \ldots)$, $q_j = (q_{1j}, q_{2j}, \ldots)$, *and* $q_k = (q_{1k}, q_{2k}, \ldots)$.

This result has several interesting consequences. First, we see that the asymptotic correlation structure is free of α, which determines the level of sparsity between groups, although the relative weights of the groups λ_g do play a role. Second, ψ is directly involved in the correlation, with the maximal correlation determined by the angle between q_j and q_k under the inner-product induced by Λ, which will often simply be the identity matrix.

Special cases of this result give further insight. Consider the non-overlapping scenario in which q_j and q_k have exactly one non-zero entry. If the predictors do not share the same group, the correlation is 0, so that predictors which do not share groups are asymptotically uncorrelated. On the other hand, if the predictors lie in the same group g then we have an asymptotic correlation of

$$\frac{\psi q_{gj} q_{gk}}{\sqrt{\psi q_{gj}^2 + q_{gj}} \cdot \sqrt{\psi q_{gk}^2 + q_{gk}}} = \frac{\psi}{\sqrt{\psi + q_{gj}^{-1}} \cdot \sqrt{\psi + q_{gk}^{-1}}}.$$

In Section 3.2 we recommend, as a possible default, setting $q_{gj} = (G_j P_g)^{-1}$ where G_j is the number of groups predictor j lies in and P_g is the (modified) size of group g. For the non-overlapping setting, this gives $\psi/(\psi + P_g)$ for the correlation. When ψ is large relative to the group size, we have a correlation of 1.

3.2. The Role of the Hyperparameters

The hyperparameters α and ψ, as well as the weights λ and Q, play a large role in the determining the properties of the OG-Dirichlet prior. The parameter α determines the *between-group* sparsity level. As $\alpha \to 0$, the Dirichlet distribution $\pi \sim \text{Dirichlet}(\alpha\lambda_1, \ldots, \alpha\lambda_G)$ degenerates to a $\text{Categorical}(\lambda_1, \ldots, \lambda_G)$ distribution, so that when α is small we will only have one active group; conversely, as $\alpha \to \infty$, the distribution of π converges to a point mass at $(\lambda_1, \ldots, \lambda_G)$, so that the groups are selected according to prior weights. The parameter ψ determines the *within-group* sparsity, with $\psi \to 0$ causing $\omega_g = (\omega_{g1}, \ldots, \omega_{gP})$ to

converge to a $\text{Categorical}(q_{g1}, \ldots, q_{gP})$ distribution, and $\psi \to \infty$ causing $\omega_g \to (q_{g1}, \ldots, q_{gP})$.

By determining the within and between group sparsity levels, α and ψ also determine the sensitivity to false positives and negatives. For large α's, we expect many groups to be relevant, inflating the probability of falsely flagging a group as relevant. Small values of α make it more difficult for groups to enter, increasing the risk of false negatives. Similarly, when ψ is large, selected groups will tend to have all of their predictors included, inflating the risk of false positives when only a small number of variables within a relevant group are active. We elect to place exponential priors on the parameters α and ψ to allow for the data to determine reasonable values for these quantities. The impact of the prior is assessed in Section 4.

There are many options for the choice of λ and Q, depending on what prior information is available about the role of the groups. There are two situations we consider. First, it may be the case that all groups are, a priori, thought to exert an equal amount of influence on the response. In this case, a reasonable choice is $\lambda = (1/G, \ldots, 1/G)^\top$. Regardless of the choice of Q, this results in predictors which are contained in *small* groups having an inflated importance. For example, if $P = 1000$, $G = 10$, and there exists a group with only one predictor, then this expresses the opinion a-priori that this predictor will account for, on average, $1/10^{\text{th}}$ of the splits in the ensemble.

Alternatively, we might wish for our prior to reflect the opinion that all predictors in the ensemble account for, on average, $1/P^{\text{th}}$ of the splits in the ensemble, i.e., $E(W)E(\pi) = (1/P)\mathbf{1}$. Additionally, we may feel that each group should have prior importance proportional to its size. For each predictor j let G_j denote the number of groups that j belongs to and, for each group g, let $P_g = \sum_{j \in g} G_j^{-1}$, where we abuse notation and let $j \in g$ denote that predictor j is in group g. We then consider $\lambda_g = P_g/P$ and, when $j \in g$, we take $q_{gj} = (G_j P_g)^{-1}$. For each j we then have $E(s_j) = \sum_g \lambda_g q_{gj} = \sum_{g: j \in g} \frac{P_g}{P} \frac{1}{G_j P_g} = P^{-1}$ as desired.

It is difficult to strike a balance between penalizing the groups and penalizing the individual predictors. Our experience is that, when λ and Q are chosen so that $E(s_j) = P^{-1}$, this results in the model favoring groups with many predictors. Alternatively, under the alternative weights with $\lambda_g = G^{-1}$, predictors which appear in many groups and/or small groups are favored. These general trends were observed by Obozinski et al. (2011), who studied a variety of issues related to selection of weights for the overlapping grouped lasso, and ultimately resolved this issue for the breast cancer dataset by only considering groups with fewer than 50 variables. In both our simulation study and for the real data, we use the above specification with $\lambda_g = P_g/P$ and $q_{gj} = (G_j P_g)^{-1}$, with the understanding that this will tend to favor large groups.

4. Simulation Study

4.1. Non-overlapping Groups

We first evaluate the OG-BART model in the non-overlapping setting. We simulated a non-overlapping group structure in the following manner. We first simulated a probability vector $V = (V_1, \ldots, V_G)$ according to a truncated stick breaking distribution in which $V_k = V'_k \prod_{j<k}(1 - V'_j)$ where $V'_{5000} = 1$ and $V'_k \overset{\text{indep}}{\sim} \text{Beta}(0.01, 0.99k)$. Here, V is a truncation of the two parameter Poisson-Dirichlet process of Pitman & Yor (1997). This generative scheme results in several large groups, but is "heavy tailed" in the sense that there are also a large number of small groups, which is typical when the groups correspond to genetic pathways. Each predictor X_j then appeared in a given group k with probability V_k. The same grouping structure was used in all simulations, although which groups were relevant varied depending on the simulation scenario.

We take $f_0(x)$ to be the function introduced by Friedman (1991) given by $f_0(x) = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5$. Rather than using the coordinates $1, \ldots, 5$, however, we select a group of size $P_g \in \{5, 50\}$. The $P_g = 5$ setting allows us to examine the situation in which every predictor in the group is relevant, so that the response exhibits sparsity between groups but not sparsity within groups. The $P_g = 50$ setting allows us to examine what happens when there is sparsity both within and between groups. We then set $Y_i = f(X_i) + \sigma\epsilon_i$ with $\epsilon_i \overset{\text{iid}}{\sim} \text{Normal}(0, 1)$. The predictors X_i were given marginal $\text{Uniform}(0, 1)$ distributions with a Gaussian copula to induce correlated within group. We set $\text{Cor}(\Phi^{-1}(X_{ij}), \Phi^{-1}(X_{ik})) = 0.63$ if X_{ij} and X_{ik} are in the same group, and 0 otherwise. We set $\lambda_g = P_g/P$ and $q_{gj} = 1/(G_j P_g)$ as described in Section 3.2. We considered a grid of σ on $[1, 10]$ and the experiment was repeated 100 times for each σ on the grid.

We monitor the number of *false positives* (FP), the number of *false negatives* (FN), and the integrated root mean squared error (RMSE) $\|f_0 - \widehat{f}\|_2$, where $\|g\|_2 = (\int g^2 \, dF_0)^{1/2}$ and F_0 denotes the true distribution of the X_i's. Additionally, we monitor the F_1 score, which we define as $\text{F}_1 = \frac{2\,\text{TP}}{2\,\text{TP} + \text{FN} + \text{FP}}$ where TP denotes the number of *true positives*. This quantity, which is the harmonic mean of the precision $\text{TP}/(\text{TP} + \text{FP})$ and the recall $\text{TP}/(\text{TP} + \text{FN})$, is a commonly used summary for how well a variable selection procedure works (see, e.g., Nan & Yang, 2014).

Competing methods: We primarily focus on the comparison between BART with the OG-Dirichlet prior (OG-BART) and BART with the Dirichlet prior (SBART). In both cases,

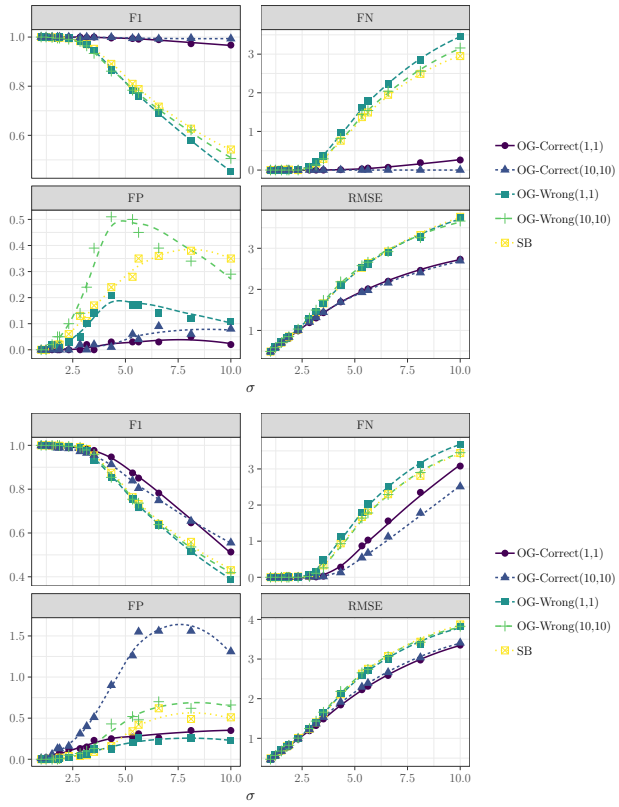

Figure 1. Results for the simulation of Section 4.1. The top panel gives results when the relevant group contains $P_g = 5$ relevant predictors, while the lower panel gives results when the relevant group contains $P_g = 50$ predictors. "SB" denotes SBART, "OG-Correct(a, b)" denotes OG-BART with the correct grouping structure and prior means a and b for α and ψ, and "OG-Wrong" denotes the use of the incorrect grouping structure.

we use soft decision trees in place of traditional decision trees to improve performance. We narrow our focus because SBART has been shown to greatly outperform the relevant competitors on this particular example (Linero & Yang, 2018). We first allow OG-BART to either have a correctly specified grouping structure or an incorrectly specified grouping structure. The incorrectly specified grouping structure is generated randomly using the same procedure used to generate the correct grouping structure. In each case, we give the hyperparameters α, ψ independent exponential priors, with means of either 1 or 10. For the SBART prior, the sparsity parameter α is given an $\text{Exponential}(1)$ prior. Finally, for the purpose of comparing with another approach which takes advantage of the grouping structure, we also considered the overlapping grouped lasso (Jacob et al., 2009) as implemented in the package `grpregOverlap`, but the results are omitted due to generally poor performance. To get a sense of the performance of the overlapping group lasso on this simulation, see Section 4.2.

Results: Results for the simulation are given in Figure 1. To aide visualization, splines were fit to the results of the simulation. As before, we set $\lambda_g = P_g/G$ and $q_{gj} = 1/(G_j P_g)$. When $P_g = 5$, we see a very large benefit to using correct grouping information, as the F_1 score is close to 1 even for small signal levels and the RMSE is small. Worse results are obtained when the grouping structure is incorrectly specified but, conveniently, they are no worse than the usual SBART model. When $P_g = 50$, the overall trends remain the same for F_1, FN, and RMSE, while FP exhibits slightly different behavior. First, we see that the correctly specified group prior, combined with the mean 10 exponential prior, results in a larger number of false positives. This behavior is quite curious, as a correct prior structure gives *worse* performance on this metric. This occurs because larger values of ψ discourage sparsity within group, which leads to the selection of predictors which are in the active group but do not influence the response. To confirm that this behavior is to be expected when there is both sparsity within and between groups, and is not a bug of our model, we study this behavior in the context of the much simpler "normal means" problem in the supplementary material.

Overall, OG-BART performs well, and can be safely used even when the grouping structure is incorrectly specified. In general, the method performs best when there is sparsity between groups and density within groups. We also observe that hyperpriors for α and ψ determine the tradeoff between false positives and false negatives as described in Section 3.2.

4.2. Overlapping Groups

We next consider the setting in which predictors are members of potentially many groups. We simulate V as before, but instead of having each predictor belong to a single group we instead sample the number of groups X_j from a $\text{Poisson}(1)$ distribution. When predictor j belongs to no groups, we take X_j to belong to a group of its own.

We draw the regression function $f(x)$ from an SBART prior $f \sim \text{SBART}$ with 50 trees, in which $s_j = 1/9$ for a subset of X_j's in a group of size 46, and $s_j = 0$ for all other j's. For each simulated dataset, we sampled $f(x)$ from the prior and generated $Y_i = f(X_i) + \sigma\epsilon_i$. As before, we set $P = 1000$ for the number of predictors and $N = 250$ for the number of observations. As in the nonoverlapping setting, we monitor the F_1, FN, FP, and RMSE statistics. We varied σ on a grid from 0.1 to 1 evenly spaced on the log scale and replicated the experiments 100 times.

Competing methods: In addition to SBART and OG-BART, we also display results for the overlapping group

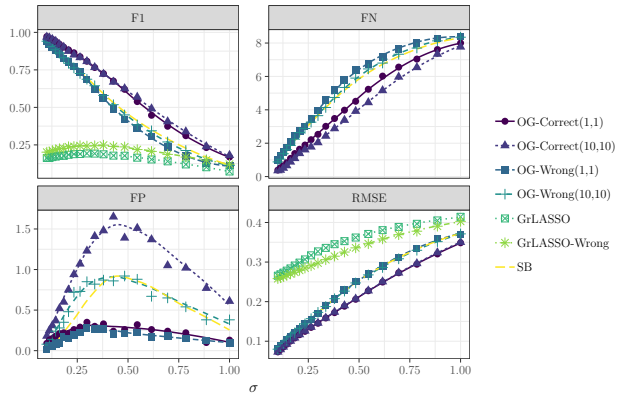

Figure 2. Simulation results for the simulation of Section 4.2, with the same conventions as Figure 1.

lasso (Jacob et al., 2009) with tuning parameters selected by cross validation. We attempted to fit variants of the sparse overlapping group lasso to allow for within-group sparsity, but were unsuccessful with publicly available software for the simulated data; for a comparison with the bi-level MCP procedure of Breheny & Huang (2009), see Section 5.

Results: Results are given in Figure 2. Speaking roughly, the trends are quite similar to what is observed in the previous simulation. When the group structure is correctly specified, we see an improvement in F_1, FN, and RMSE, while when the group structure is not correctly specified we do not pay any price. The overlapping group lasso performs extremely poorly across all metrics, and was omitted from the FN and FP plots due to dominating the figures. Part of the reason for this poor performance is that the overlapping group lasso does not perform within-group selection. We again observe that larger values of $E(\alpha)$ and $E(\psi)$ result in larger false positive rates. This experiment also considers substantially lower signal levels than the previous one, with the F_1 score decaying nearly to 0 for larger values of σ.

5. Application to Breast Cancer Data

We now illustrate the use of OG-BART on the gene expression dataset of Van De Vijver et al. (2002). This dataset consists of gene expression data on $8,141$ genes and a total of $N = 295$ breast cancer tumors, of which 217 were non-metastatic and 78 were metastatic. We use a preprocessed version of this dataset from Jacob et al. (2009) in which groups are formed from a subset of the canonical pathways of MSigDB (Subramanian et al., 2005), giving $G = 637$ pathways. The data is further subset by considering only genes which are included in these pathways, giving a subset of $P = 3510$ genes. We use the data augmentation strategy described in Section 2.3 to fit a probit model to the probability of a tumor being metastatic.

We fit the model using several different settings for the hyperparameters and hyperpriors for α and ψ, with the weights $\lambda_g = P_G/P$ and $q_{gj} = (G_j P_g)^{-1}$ described in Section 3.2. We report results for the setting $\alpha \sim \mathrm{Exp}(1)$ and $\psi = 1$, as this leads to sparser/more interpretable models. A comparison of the predictive performance across different hyperparameter settings suggests that somewhat larger values, say $\alpha, \psi \sim \mathrm{Exp}(10)$, leads to better predictive performance, but typically includes a larger number of predictors and groups.

Before proceeding, we remark that there is not overwhelming evidence for any particular gene being active, in the sense that there are many disjoint sparse models which predict the response well. This is caused by the combination of (a) a massive number of genes and pathways (many of which are marginally correlated with the response) with (b) a prior which encourages a high degree of within and between group sparsity and (c) a small sample size. With that in mind, the model flags the gene TK1 (thymidine kinase 1) as having a relatively large probability ($\approx 57\%$) of being included in the model. Thymidine kinase serum levels are known to be elevated in progressive breast cancers (Topolcan & Holubec Jr, 2008). The model also suggests the relevance of TXNRD1 (thioredoxin reductase 1, $\approx 25\%$), which is also known to be associated with prognosis of breast cancer. Suppression of thioredoxin reductase enzymes has been suggested as a promising avenue for anti-cancer treatments (Cadenas et al., 2010).

We now use the posterior of OG-BART to analyze the information in the genetic pathways. We consider a pathway g as active if $Z_b = g$ for some branch b in the ensemble. The top two pathways selected by OG-BART both account for six of the top ten genes selected in the model, including TK1 and TXNRD1. As these two pathways overlap considerably (overlap coefficient 85%), we discuss them jointly. The marginal probability of at least one of these two pathways being active was 62%. To get a sense of why these pathways were selected, we give in Figure 3 a histogram of the P-values obtained by a Wilcoxon signed rank test comparing the values of each predictor for $Y_i = 0$ and $Y_i = 1$. We see that this pathway contains many genes which are differentially expressed in metastatic tumors. Consequently, the model favors the inclusion of this pathway.

Figure 3 also includes the third most commonly occurring pathway in the ensemble. Interestingly, while there is no strong evidence of any *single* gene being active, the model still includes this pathway frequently. This pathway consists primarily of genes coding for ribosomal proteins, which are thought to play an important role in cancer development (Goudarzi & Lindström, 2016). The P-values from the Wilcoxon signed ranked tests show that many of the genes

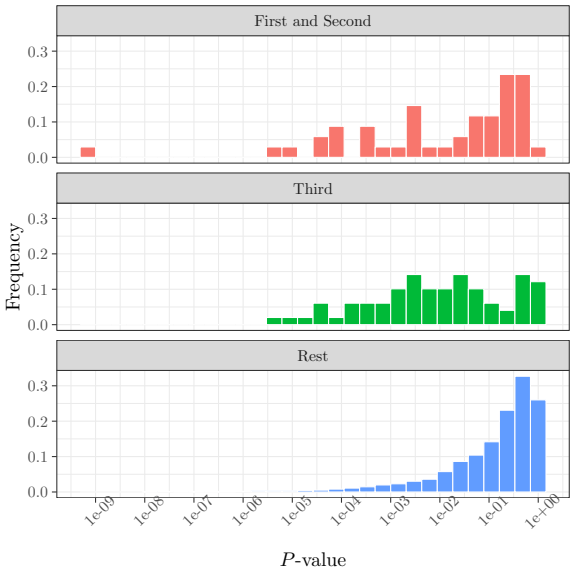

Figure 3. Histogram of P-values from Wilcoxon's signed rank test for each predictor in the first and second most frequently occurring pathways (top), the third most frequently occurring pathway (middle), and all remaining pathways (bottom).

in this pathway are marginally correlated with the response; in fact, it has a larger frequency of P-values below 10^{-2} than the first two pathways. Due to correlation between the genes in this pathway, however, it is difficult for any single gene from this pathway to stand out when analyzed jointly with the SBART model.

Next, we give a sanity check for OG-BART, and demonstrate the need for methods which allow for non-linearities, by assessing the predictive performance of the overlapping group lasso, SBART, and OG-BART. Additionally, we consider the bi-level MCP selection procedure (cMCP) of Breheny & Huang (2009), which like OG-BART allows for selection both within and between groups. We perform 5-fold cross validation, replicated 5 times, and compute the heldout deviance $D = -2\sum_{i=1}^{N}[Y_i \log \widehat{\pi}_i + (1 - Y_i)\log(1 - \widehat{\pi}_i)]$. Following Jacob et al. (2009), we balance the data by adding two replicates (three in total) for each metastatic tumor (keeping all replicates in the same fold during cross-validation). Results for each method are given in Table 1. First, we see that the overlapping group lasso performs relatively poorly, giving some evidence for the need for methods which allow non-linearities; additionally, the fact that cMCP outperforms the overlapping group lasso suggests that taking performing group-level selection alone is insufficient to obtain good performance. Second, we see that the OG-BART performs somewhat better than SBART, with OG-BART also outperforming SBART on all five splits.

Method	Average Heldout Deviance	
OG-BART	620	
SBART	646	(0.005)
OG-Lasso	797	(< 0.0001)
cMCP	698	(0.014)

Table 1. Average deviance on held-out data computed using 5 replications of 5-fold cross validation; parentheses gives the P-value obtained from a paired t-test comparing to OG-BART.

6. Discussion

In this paper, we incorporated grouping information into nonparametric prediction and variable selection tasks using the OG-Dirichlet prior with Bayesian additive regression trees. While we have developed the methodology in a manner specific to BART, we believe that these methods should also be applicable to greedy decision-tree construction algorithms such as boosting by using the prior as a penalization term. Such extensions could provide large computational benefits. Additionally, we have used Markov chain Monte Carlo to fit our models; an attractive alternative, which scales to large datasets, is the accelerated Bayesian additive regression trees framework recently proposed by He et al. (2019), which allow for BART models to be fit in time comparable to the popular `xgboost` package. Using such modifications could allow for applications to much larger data than we have considered here.

This paper has focused only on prior information in the form of groups. Other forms of prior information will be examined in later works. In the case of genetic pathways, we have only taken into account whether a gene is in a specific pathway or not. This is not the only information available, however, as genetic pathways may also have a graphical structure in which the genes correspond to vertices of the graph. Use of network structure has been observed to further improve the performance of variable selection methods for linear models (Chang et al., 2016; Li & Li, 2008). While Dirichlet type priors are well-suited to the grouping structures studied here, they are not well-suited to encoding structures expressed by general undirected graphs.

References

Albert, J. H. and Chib, S. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88:669–679, 1993.

Basu, S., Kumbier, K., Brown, J. B., and Yu, B. Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences*, pp. 201711236, 2018.

Blei, D. and Lafferty, J. Correlated topic models. *Advances*

in Neural Information Processing Systems, 18:147, 2006.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3: 993–1022, 2003.

Breheny, P. and Huang, J. Penalized methods for bi-level variable selection. *Statistics and its interface*, 2(3):369, 2009.

Breiman, L. Random forests. *Machine Learning*, 45(1): 5–32, 2001.

Cadenas, C., Franckenstein, D., Schmidt, M., Gehrmann, M., Hermes, M., Geppert, B., Schormann, W., Maccoux, L. J., Schug, M., Schumann, A., et al. Role of thioredoxin reductase 1 and thioredoxin interacting protein in prognosis of breast cancer. *Breast cancer research*, 12(3): R44, 2010.

Candes, E. and Tao, T. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

Chang, C., Kundu, S., and Long, Q. Scalable Bayesian variable selection for structured high-dimensional data. *Biometrics*, 2016. To appear.

Chipman, H., George, E. I., Gramacy, R. B., and McCulloch, R. Bayesian treed response surface models. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(4):298–305, 2013.

Chipman, H. A., George, E. I., and McCulloch, R. E. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

Denison, D. G., Mallick, B. K., and Smith, A. F. A Bayesian CART algorithm. *Biometrika*, 85(2):363–377, 1998.

Dennis, S. Y. On the hyper-Dirichlet type 1 and hyper-Liouville distributions. *Communications in Statistics - Theory and Methods*, 20(12):4069–4081, 1991.

Dorie, V., Hill, J., Shalit, U., Scott, M., and Cervone, D. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *arXiv preprint arXiv:1707.02641*, 2017.

Du, J. and Linero, A. R. Interaction detection with bayesian decision tree ensembles. In *22nd Proceedings of the International Conference on Artificial Intelligence in Statistics (AISTATS)*, 2019.

Freund, Y., Schapire, R., and Abe, N. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 4(5):771–780, 1999.

Friedman, J. H. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67, 1991.

Goudarzi, K. M. and Lindström, M. S. Role of ribosomal protein mutations in tumor development. *International journal of oncology*, 48(4):1313–1324, 2016.

He, J., Yalov, S., and Hahn, P. R. Accelerated Bayesian Additive Regression Trees. In *22nd Proceedings of the International Conference on Artificial Intelligence in Statistics (AISTATS)*, 2019.

Irsoy, O., Yildiz, O. T., and Alpaydin, E. Soft decision trees. In *Proceedings of the International Conference on Pattern Recognition*, 2012.

Jacob, L., Obozinski, G., and Vert, J.-P. Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*, pp. 433–440. ACM, 2009.

Kanehisa, M. and Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

Kapelner, A. and Bleich, J. bartMachine: Machine learning with Bayesian additive regression trees. *Journal of Statistical Software*, 70(4):1–40, 2016.

Lakshminarayanan, B., Roy, D. M., and Teh, Y. W. Mondrian forests: Efficient online random forests. In *Advances in Neural Information Processing Systems*, pp. 3140–3148, 2014.

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

Li, F. and Zhang, N. R. Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American statistical association*, 105(491):1202–1214, 2010.

Linero, A. R. A review of tree-based bayesian methods. *Communications for Statistical Applications and Methods*, 24(6):543–559, 2017.

Linero, A. R. Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association*, 113(522):626–636, 2018.

Linero, A. R. and Yang, Y. Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5):1087–1110, 2018.

Linero, A. R., Sinha, D., and Lipsitz, S. R. Semiparametric Mixed-Scale Models Using Shared Bayesian Forests. *arXiv e-prints arXiv:1809.08521*, 2018.

Minka, T. The Dirichlet-tree distribution. Unpublished manuscript available at research.microsoft.com, 1999.

Murray, J. S. Log-linear Bayesian additive regression trees for categorical and count responses. *arXiv preprint arXiv:1701.01503*, 2017.

Nan, Y. and Yang, Y. Variable selection diagnostics measures for high-dimensional regression. *Journal of Computational and Graphical Statistics*, 23(3):636–656, 2014. doi: 10.1080/10618600.2013.829780.

Neal, R. M. *Bayesian Learning For Neural Networks*. PhD thesis, University of Toronto, 1995.

Obozinski, G., Jacob, L., and Vert, J.-P. Group lasso with overlaps: the latent group lasso approach. *arXiv preprint arXiv:1110.0413*, 2011.

Pitman, J. and Yor, M. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900, 04 1997.

Pratola, M., Chipman, H., George, E., and McCulloch, R. Heteroscedastic BART using multiplicative regression trees. *arXiv preprint arXiv:1709.07542*, 2017.

Rockova, V. and Lesaffre, E. Incorporating grouping information in Bayesian variable selection with applications in genomics. *Bayesian Analysis*, 9(1):221–258, 03 2014.

Rockova, V. and van der Pas, S. Posterior concentration for Bayesian regression trees and their ensembles. *arXiv preprint arXiv:1078.08734*, 2017.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

Sparapani, R. A., Logan, B. R., McCulloch, R. E., and Laud, P. W. Nonparametric survival analysis using Bayesian additive regression trees (BART). *Statistics in medicine*, 2016.

Starling, J. E., Murray, J. S., Carvalho, C. M., Bukowski, R., and Scott, J. G. Functional response regression with funbart: an analysis of patient-specific stillbirth risk. *arXiv preprint arXiv:1805.07656*, 2018.

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

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43): 15545–15550, 2005.

Tibshirani, R. J. and Taylor, J. The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335–1371, 06 2011.

Topolcan, O. and Holubec Jr, L. The role of thymidine kinase in cancer diseases. *Expert opinion on medical diagnostics*, 2(2):129–141, 2008.

Van De Vijver, M. J., He, Y. D., Van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., and Marton, M. J. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25):1999–2009, 2002.

Wang, L., Chen, G., and Li, H. Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics*, 23(12):1486–1494, 2007.

Xu, X. and Ghosh, M. Bayesian variable selection and estimation for group lasso. *Bayesian Analysis*, 10(4): 909–936, 2015.

Yang, Y. and Dunson, D. B. Minimax optimal Bayesian aggregation. *arXiv preprint arXiv:1403.1345*, 2014.

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

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.