Modelling Heterogeneity Using Bayesian Structured Sparsity

Max Goplerud*

Abstract

How to estimate heterogeneity, e.g. the effect of some variable differing across observations, is a key question in political science. Methods for doing so make simplifying assumptions about the underlying nature of the heterogeneity to draw reliable inferences. This paper allows a common way of simplifying complex phenomenon (placing observations with similar effects into discrete groups) to be integrated into regression analysis. The framework allows researchers to (i) use their prior knowledge to guide which groups are permissible and (ii) appropriately quantify uncertainty. The paper does this by extending work on "structured sparsity" from a traditional penalized likelihood approach to a Bayesian one by deriving new theoretical results and inferential techniques. It shows that this method outperforms state-of-the-art methods for estimating heterogeneous effects when the underlying heterogeneity is grouped and more effectively identifies groups of observations with different effects in observational data.

Pages: 34 Pages
Key Words: structured sparsity; generalized LASSO; Bayesian methods; heterogeneous effects; machine learning

1 Introduction

Data analysis in social science faces a fundamental trade-off. On the one hand, there is a huge amount of richness and complexity in each observation. Yet, statistical models abstract away from this by making assumptions regarding the comparability of observations, the constant effects of some variables, functional form assumptions, and more. The choice of assumptions is not merely a technical decision but can have crucial substantive implications. Different choices may lead to different substantive conclusions, and all sets of assumptions imply particular relationships between the observations. Whenever possible, the choice of assumptions should be motivated by the researcher's substantive knowledge of the underlying question.

Consider a standard regression context. Perhaps the most fundamental question is to understand the effect that some independent variable has on the outcome. In doing so, the most common—and strongest—assumption would be that this variable has the same effect on all observations or that a single aggregate effect was substantively interesting. Straight-forwardly including the variable linearly in the regression provides an estimate of this effect, as long as other critical assumptions are satisfied. Focusing on that quantity, however, may mask important sub-group heterogeneity that the researcher wishes to explore. A much more flexible approach to doing so would be to estimate a separate effect for each combination of the other covariates. Unfortunately, this approach typically leads to unacceptably noisy estimates of the sub-group effects. Research into simplifying the problem by assuming some stability in the underlying heterogeneity is often known as "estimating heterogeneous treatment effects" (e.g. Imai and Strauss 2011; Grimmer, Messing, and Westwood 2017).

Existing approaches to simplification often bet on the idea that the "no heterogeneity" assumption (i.e. a single effect) is approximately right: Estimates for sub-groups should be stabilized by being pulled towards some global or aggregate effect. This assumption appears in hierarchical models, sparse methods, and—more implicitly—many methods based on regression trees. A key goal of this paper is to suggest, however, that this assumption,

while useful in many circumstances, does not always match how researchers intuitively and naturally make sense of a complex phenomenon.

Rather, this paper builds on a different intuition: Complex phenomena can be understood by classifying observations into a small number of *groups* while being guided by our prior knowledge (if available). Such group creation is ubiquitous in political science research and appears whenever scholars create typologies or categorical variables. Explanations based on groups are often desirable because they also easily interpretable and explainable to others. Further, researchers are often in the position of having *some* prior knowledge or belief about group structure (e.g. neighbors are connected), although they are not sufficiently confident to authoritatively classify observations. Thus, traditional methods of encoding and estimating groups (e.g. indicator variables or hierarchical models that allow for heterogeneity by group) are unsuitable as a core assumption of those approaches is that the groups are known *ex ante*. On the other extreme, existing clustering algorithms are often unable to effectively integrate the researcher's prior beliefs on group membership and thus fail to effectively leverage valuable information from prior research.

This paper provides a framework to quantitatively encode this intuition: At its core is an existing method known as "structured sparsity" (Huang, Zhang, and Metaxas 2011; Bach et al. 2012; Chen et al. 2012). It starts from an agnostic scenario where there are many parameters representing the effect in each small unit (e.g. treatment-covariate profile combination). Before estimating the model, the researcher then decides which of these units *might* be directly connected together in the same group—e.g. neighboring counties or profiles with common characteristics—based on their prior knowledge of the substantive question at hand. Finally, the model is estimated and the data determines whether two units should be fused together (i.e. given the *same* effect). This occurs if (a) their estimated effects are close and (b) prior knowledge permits their combination. Groups thus emerge from this process where two units are given the same effect. The name "structured sparsity" comes from the fact that it modifies a traditional sparse approach (e.g. the LASSO; Tibshirani 1996) in the

following crucial way: While sparsity seeks parsimony by encouraging many parameters to be *zero* ("sparse"), structured sparsity encourages many parameters to be *equal*—resulting in clustering.

Unfortunately, in its existing form, structured sparsity is not suitable for most social scientific research. There are two main limitations. First, it is difficult to quantify uncertainty in the estimated parameters as standard techniques of differentiating the objective or bootstrapping are not appropriate. Second, existing inferential techniques are limited to particular structures and/or a linear likelihood. This leaves an undesirable situation where researchers may think that particular choices of structure *should* be used, but would be unable to fit a model that matches their theoretical belief.

I address these concerns by creating a Bayesian formulation of structured sparsity. This allows simple inference and uncertainty quantification for many likelihoods and any set of prior beliefs (structure). In doing so, I generalize existing Bayesian research that is focused on particular structures (e.g. Kyung et al. 2010; Betancourt, Rodríguez, and Boyd 2017; Tansey et al. 2017; Faulkner and Minin 2018). The prior proposed to induce Bayesian structured sparsity is novel, and thus I provide new theoretical results on when the resulting posterior is proper as the prior is usually improper by design. Further, since the non-Bayesian structured sparse estimate is of special interest, the second major technical contribution of the paper provides a new algorithm for fast estimation using an Expectation Maximization algorithm (Dempster, Laird, and Rubin 1977). This generalizes non-Bayesian estimation techniques that are also limited to specific structures or likelihoods (e.g. Arnold and Tibshirani 2016; Zhu 2017; Chen et al. 2012; Tansey et al. 2017).

The framework outlined in this paper is most applicable in the common scenario where the researcher has a large number of categorical or binary variables and wishes to simplify the complexity by creating groups. I explore two different settings to show its effectiveness: First, I use simulations to sharpen the intuition that existing methods perform poorly when the underlying heterogeneity is based around unknown *groups*. Using a simple example

where half of the units have an effect of one, half have an effect of negative one, but their membership is unknown, I show that many state-of-the-art methods (e.g. BART, LASSO, etc.) perform poorly—even losing to conventional methods such random effects. Structured sparsity outperforms both sets of methods, however, as it pools information more effectively by *creating* groups of observations with the same effect. This creates both better performing and more easily interpretable results versus conventional methods.

Second, I re-examine an experiment on the effect of politicians claiming credit for spending projects in their districts (Grimmer, Westwood, and Messing 2014). Existing work has re-analyzed this study to look for heterogeneous effects across the types of treatment and respondent (Grimmer, Messing, and Westwood 2017), and it provides a fruitful test of how structured sparsity's intuition based on groups compares to existing approaches.

Substantively, I show that this method resolves a puzzling result in the initial analysis where (i) conservatives reacted more *positively* to Planned Parenthood centers than moderates and (ii) conservatives reacted more *negatively* to the construction of gun ranges than moderates for certain combinations of treatment effects. Further, I demonstrate that the preference for a project sponsored by a co-partisan (Grimmer, Westwood, and Messing 2014) is concentrated amongst conservative Republicans and, to a lesser degree, liberal Democrats. I further provide evidence of the importance of the role of groups by putting structured sparsity into an ensemble with many state-of-the-art methods. It show that it gets substantial non-zero weight and is often the highest weighted model. This suggests that relying on groups captures something important and distinct from existing approaches.

2 Heterogeneous Effects of Credit-Claiming

I begin with a motivating example about legislator credit-claiming by Grimmer and co-authors (Grimmer, Westwood, and Messing 2014; Grimmer, Messing, and Westwood 2017). It experimentally tests a long-standing theory in American politics where voters are believed

to increase support for their elected representatives if their representative can "claim credit" for projects in their district such as building a road or a bridge (Mayhew 1974). I specifically focus on an experiment that explores which types of projects may get more "reward" from voters and whether there are certain types of projects that voters dislike. Table 1 outlines the factorial-style experiment fielded on Mechanical Turk where voters are either shown a control message about no project (10% of the sample) or a hypothetical project (Grimmer, Westwood, and Messing 2014, pp. 97-105; Grimmer, Messing, and Westwood 2017). It yields 216 unique treatments.

Table 1: Credit-Claiming Experiment

(a) Description of Treatment Factors

- Type: Roads, Police, Parks, Fire Department, Gun Range, Planned Parenthood

- Money: $50 thousand or $20 million

- Stage : Will Request, Requested, or Secured

- Sponsor: Republican or Democrat

- Co-Sponsor: None, Senate Republican, or Senate Democrat

(b) Sample Treatment

- Representative X, Democrat, and Senator Y, a Republican, requested $20 million to purchase safety equipment for local police. Rep. X said "This money would help our brave police officers stay safe as they protect our property from criminals."

- Representative Z, Democrat, secured $50 thousand for medical equipment at the local planned parenthood. Rep. Z said "This money will help provide state of the art care for women in our community."

Note: A brief description of each of the treatment factors is found above. Each treatment consists of one level of each of the factors combined together into a message. Two examples are shown on the right. The full text of each treatment can be found in Grimmer, Westwood, and Messing (2014, p. 99) or in Grimmer, Messing, and Westwood (2017).

After having seen the project, the voters are asked whether "they 'approve or disapprove' of the way the fictitious representative 'is performing (his/her) job in Congress'" (Grimmer, Westwood, and Messing 2014, p. 98); the outcome is binarized for their analysis. Grimmer, Messing, and Westwood (2017) consider whether the effects of a particular treatment vary by respondent characteristics; for example, do Democrats prefer certain types of projects more than Republicans? They also explore whether certain aspects of the treatment have

interactive effects.

Such an analysis of heterogeneous effects can be conceptualized in the following way. First, the researcher creates an "expanded" design matrix that includes many possible interactions between treatments and respondents. Grimmer, Messing, and Westwood (2017) include all pairwise interactions between the treatment factors (e.g. type-money, type-stage, money-stage, etc.) as well as interacting those pairwise interactions with respondent characteristics (e.g. ideology-type-money). This generates a problem for inference as there are only 1,074 responses but nearly 300 covariates in the design mentioned above. Thus, some sort of stabilization is needed to draw reliable inferences.

Existing approaches to heterogeneous effects typically propose methods that make some variant of the following assumption: the majority of deviations from an "aggregate" effect should be stabilized by pulling them towards zero. That is, they presume a model where a single global effect (i.e. "no heterogeneity") can well describe most observations and that deviations from that global effect are likely small for most units. This characterizes hierarchical models with random slopes (Gelman and Hill 2006), sparse models (Imai and Ratkovic 2013), and many methods based on regression trees (e.g. BART, causal forests; Hill 2011; Wager and Athey 2018).[1]

However, this paper suggests that a different approach based on *groups* may be more fruitful. As the variables are categorical or binary (as is common in social science), a natural way to think about simplifying the estimates is creating groups of treatments with similar effects. For example, the main effect of roads and parks might be grouped and/or the interactions between party and fire departments may be fused. As the number of groups and their membership is unknown (unlike in hierarchical models), a key benefit of structured sparsity is to *estimate* the groups in a data-driven but theory-guided way.

---

1. Consider a tree placed on a set of parameters corresponding to a one-hot coding of membership into units u. A tree is grown by first deciding which unit u should be pulled off into its own group. The next step decides whether a second unit u' should form it own group. And so on. As the complexity of the tree is limited to prevent over-fitting, only a small number of units will be given a heterogeneous effect versus a global baseline containing most units. See Appendix D for more discussion.

3 Modelling Heterogeneity Using Structured Sparsity

I focus in the remainder of the paper on structured sparsity as it is a flexible method for creating groups that can explicitly incorporate theoretical information about group structure (Huang, Zhang, and Metaxas 2011; Bach et al. 2012; Chen et al. 2012). I outline the approach in the context of a simplified version of the credit-claiming experiment with two treatment factors: project type (Gun Range ["Gun"], Planned Parenthood ["PP"], or Road) and amount of funding ($50k or $20million).[2]
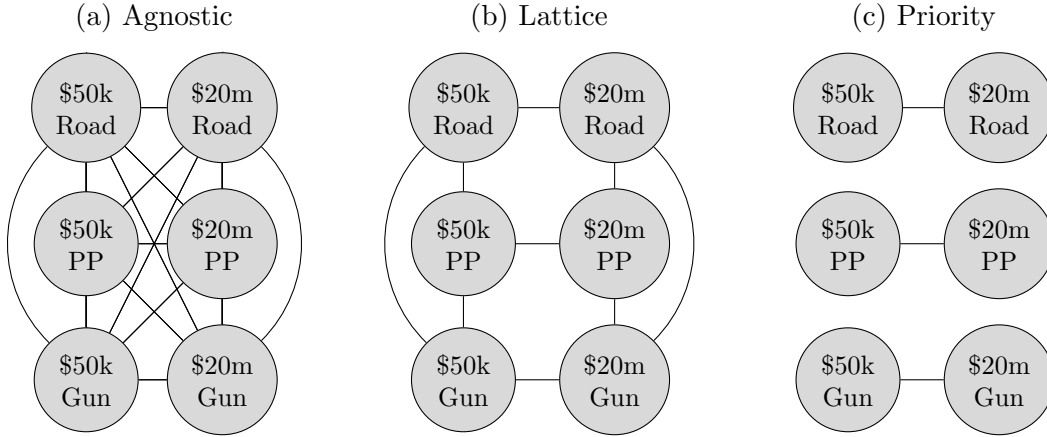
The basic procedure creates a large number of interactions defining a set of units with the same effect. Groups of units are fused together and given the same effect and thereby clusters emerge. Thus, unlike "top-down" clustering approaches such as finite mixture models, groups emerge from the (pairwise) comparison of units and thus it is a "bottom-up" approach to creating clusters.

The researcher's prior knowledge is critical here in deciding which comparisons are permissible. It is often the case that researchers wish to rule out certain units being put together and thus wish to constrain the group formation process. These beliefs about which units are possibly connected together gives the "structure" to structured sparsity. Structure is often helpfully illustrated visually; Figure 1 begins by creating a network of the units or effects under consideration (project type / amount combinations) and then linking together nodes that *might* be connected together.

Figure 1 shows three possible structures for this application. First, there is a fully connected or "agnostic" structure (1a); it encodes the minimal possible prior knowledge in that any two effects could be clustered together and thus any groups are possible. While it is the most flexible, it does not include any prior knowledge; it therefore may be less efficient

---

2. Alternative approaches based on clustering methods (e.g. Bonhomme and Manresa 2015; Shen and He 2015; Shahn and Madigan 2017; Shiraito 2016) are useful in certain contexts but lack the ability to easily incorporate prior knowledge on which groups are permissible. For example, creating only geographically contiguous groups in a spatial setting is difficult for the cited methods, but trivial for structured sparsity. A secondary limitation is that most approaches require fixing the number of groups for any single run of the model.

Figure 1: Different Types of Structured Sparsity

(a) Agnostic   (b) Lattice   (c) Priority

Note: Each figure shows a possible structure. The lines connecting each variable represent the possibility of exactly fusing the associated variables together when using structured sparsity. The left panel shows a fully connected structure (all groups are possible); the middle shows a "lattice" structure where two nodes are connected if and only if they share either project type or funding amount. The right panel show a structure where nodes are connected if they share project type.

than methods that incorporate structure. Further, it may form groups that are hard to interpret substantively given that there is no prior knowledge providing coherence to the groups.

Some substantive knowledge can be added in multiple ways; a more permissive but still flexible structure resembles a "lattice". This would suggest that two units can be fused together if they share *either* type of project *or* amount. Figure 1b shows the visual representation. Note that two units might still be in the same group if they are connected indirectly; for example, Roads-$50k and Roads-$20million are fused and Roads-$20million and Gun Range-$20million are fused together. Thus, Roads-$50k and Gun Range-$20 million are in the same group although they do not share any similar characteristics. It does, however, ensure that no group can contain both of them without an appropriate connecting pathway. This gives some coherence to the groups that emerge and may make them more interpretable.

Finally, a quite restrictive structure is shown in the right panel (Figure 1c). It allows only

for groups *within* one dimension of the heterogeneity, e.g. that groups can only be formed by fusing together units with the same type; it thus "prioritizes" a particular dimension of heterogeneity. These structures are appropriate when there is a strong prior belief that one dimension of heterogeneity dominates and the other is secondary.

They also differ in their limiting case; if all connections bind for agnostic or lattice structures, then the model becomes one with a single effect—a model with no heterogeneous effects and a single treatment indicator. For the priority structure, by contrast, the limiting case is a model with one effect for each project type and thus equivalent to a model where only the factor for project type was included.

Formally, existing methods for structured sparsity operate using penalized maximum likelihood. They begin by creating a vector of coefficients ($\boldsymbol{\beta}$) that represent the effect for each interactive combination; if there are p effects, then $\boldsymbol{\beta}$ has p levels where β_i represents an indicator variable for observations that have treatment combination i (e.g. Roads-$50k). The un-penalized model can be estimated in the usual way, given a likelihood function ℓ that depends on the observed data $\boldsymbol{X}$ and outcomes $\boldsymbol{y}$.

$$\hat{\boldsymbol{\beta}}_{\mathrm{MLE}} = \arg\max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y}) \tag{1}$$

Structured sparsity is induced by adding a penalty on $\boldsymbol{\beta}$ to Equation 1 that encourages elements of $\boldsymbol{\beta}$ to be set equal to each other (e.g. Bondell and Reich 2009; Tibshirani and Taylor 2011; Gertheiss and Tutz 2010; Ma and Huang 2017; Tansey et al. 2017). To make this happen, the penalty must have two properties; following work on (regular) sparsity inducing penalties (Fan and Li 2001), it should be zero (provide no penalty) if and only if the two elements are equal and it should be non-differentiable around the point where the two elements are equal. The simplest penalty to do this is based on the LASSO (Tibshirani 1996). The LASSO is often used to set coefficients equal to zero, i.e. a penalized maximum likelihood estimate where many elements of $\boldsymbol{\beta}$ were set exactly to zero.

By contrast, *structured* sparsity notes that the goal is not to set coefficients equal to zero,

9

but rather equal to *each other* to create clusters of distinct values. This intuition is formalized by placing a penalty on the difference or gap between two coefficients: $|\beta_i - \beta_j|$. Thus, one can think of the problem as deciding which differences to set (exactly) to zero; if two coefficients have no difference between them, they are fused together into a group. More generally, this form of structured sparsity can be induced by additively placing penalties on linear combinations of some element of $\boldsymbol{\beta}$. If $\boldsymbol{d}_k \in \mathbb{R}^p$, then a penalty that includes $|\boldsymbol{d}_k^T \boldsymbol{\beta}|$ encourages $\boldsymbol{d}_k^T \boldsymbol{\beta} = 0$ to hold at the optimum.

In the case analyzed above, $\boldsymbol{d}_k$ would have one element equal to one (i.e. corresponding to β_i) and one corresponding to negative one (i.e. corresponding to β_j) and thus $\boldsymbol{d}_k^T \boldsymbol{\beta} = \beta_i - \beta_j$. For example, $|\beta_{\text{Road/50k}} - \beta_{\text{Road/20mil}}|$ would be used in all three structures as it encourages those two effects to be fused but $|\beta_{\text{Road/50k}} - \beta_{\text{Gun/20mil}}|$ would only appear in the agnostic structure as the two effects share neither attribute in common. By adding one restriction for each edge in Figure 1, one can formalize each structure with a penalty that is the sum of the absolute value of the relevant restrictions, e.g. $|\boldsymbol{d}_1^T \boldsymbol{\beta}| + |\boldsymbol{d}_2^T \boldsymbol{\beta}| + \cdots$. The specific choice of $\boldsymbol{d}_k$ to create groups of coefficients is thus only one application of structured sparsity.

A natural question is whether it is possible to cluster *multiple* effects together—rather than merely pairs. In the case of the credit-claiming example, this could fuse together levels from one factor (e.g. Roads and Guns) if and only if their interactions with other factors are all close in value. This can be done by adding a penalty of the form $\sqrt{\boldsymbol{\beta}^T \boldsymbol{F}_\ell \boldsymbol{\beta}}$ with a careful choice of symmetric positive semi-definite $\boldsymbol{F}_\ell$ that generalizes existing work on the (overlapping) group LASSO (Yuan and Lin 2006; Kyung et al. 2010; Jacob, Obozinski, and Vert 2009).

Combining these two types of penalty leads to the general definition of structured sparsity in Equation 2. It contains K linear penalties and L quadratic penalties that represent the researcher's prior belief about the structure of the underlying groups.

$$\hat{\boldsymbol{\beta}}_{\text{SSparse}} = \arg\max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y}) - \lambda \left[\sum_{k=1}^{K} |\boldsymbol{d}_k^T \boldsymbol{\beta}| + \sum_{\ell=1}^{L} \sqrt{\boldsymbol{\beta}^T \boldsymbol{F}_\ell \boldsymbol{\beta}} \right] \tag{2}$$

$\hat{\boldsymbol{\beta}}_{\text{SSparse}}$ is, unfortunately, rather difficult to estimate as the penalty is not differentiable, and standard optimization methods will fail (Tseng 2001). Many different solutions have been proposed to solve this (e.g. Chen et al. 2012; Arnold and Tibshirani 2016; Zhu 2017; Gaines, Kim, and Zhou 2018; Chen et al. 2012; Tansey et al. 2017). The cited methods are limited, however, in that it does not appear that an existing method can handle (i) non-linear models and (ii) an arbitrary choice of structure (i.e. arbitrary $\boldsymbol{d}_k$ and $\boldsymbol{F}_\ell$).

Further difficulties arise in quantifying the uncertainty around $\hat{\boldsymbol{\beta}}_{\text{SSparse}}$. This occurs because the non-differentiable penalty that induces sparsity complicates traditional methods for quantifying uncertainty. Both directly examining the Hessian of the (penalized) log-likelihood and the bootstrap are inappropriate in this setting. The non-differentiability of the penalty precludes the former and the sparsity of the estimates limits the latter as it is known that the bootstrap will fail to correctly quantify uncertainty in the standard sparse case (e.g. Leeb and Pötscher 2005; Kyung et al. 2010) and thus similar problems likely arise here.[3]

4 Bayesian Structured Sparsity

This paper proposes a different way to perform inference on $\boldsymbol{\beta}$ to resolve these problems: Bayesian inference. Existing research has considered particular choices of structure (e.g. Park and Casella 2008; Kyung et al. 2010; Betancourt, Rodríguez, and Boyd 2017; Tansey et al. 2017; Faulkner and Minin 2018; Pauger and Wagner 2019), but the general case with multiple linear and quadratic restrictions remains unexplored. Thus, it is important to analyze the theoretical properties of Bayesian structured sparsity as it appears to be a new prior in its general form.

This theoretical section has two parts; first, I explicitly derive conditions when the structured sparse prior and resulting posterior is proper. Second, I outline how the posterior can

---

3. Adapting other methods from machine learning (e.g. Chernozhukov et al. 2018) to quantify uncertainty is left for future research.

be sampled and how this leads to a novel (non-Bayesian) method for finding the solution to Equation 2. The Supporting Information provides a number of additional theoretical results. Section A.4 outlines an extension to other methods of inducing sparsity ("global-local" priors; Polson and J. G. Scott 2011); Section B provides a discussion of how to choose the optimal regularization strength (λ) in a computationally efficient manner.

4.1 Theoretical Analysis of Bayesian Structured Sparsity

Existing Bayesian work using particular structures does not examine conditions for posterior propriety. Most applications note the prior is improper and adjust it in some *ad hoc* fashion to ensure propriety (e.g. Betancourt, Rodríguez, and Boyd 2017; Faulkner and Minin 2018). This is an important issue to rectify as an improper posterior could result from an improper prior. As Bayesian algorithms will seemingly function properly in this case where the model is not well-defined (Hobert and Casella 1996), this poses a danger to reliable inference. This section resolves the problem by deriving some testable conditions for posterior propriety. To begin, Equation 2 is formalized as a prior in Definition 1.

Definition 1 *A "structured sparse" prior on $\boldsymbol{\beta} \in \mathbb{R}^p$ with some regularization strength $\lambda > 0$ penalizes K linear constraints ($\boldsymbol{d}_k$) and L quadratic constraints ($\boldsymbol{F}_\ell$) on the parameters where $\boldsymbol{F}_\ell$ is symmetric and positive semi-definite. The kernel of the prior can be expressed as follows:*

$$p(\boldsymbol{\beta}) \propto \exp\left(-\lambda \left[\sum_{k=1}^{K} |\boldsymbol{d}_k^T \boldsymbol{\beta}| + \sum_{\ell=1}^{L} \sqrt{\boldsymbol{\beta}^T \boldsymbol{F}_\ell \boldsymbol{\beta}}\right]\right)$$

Further define $\boldsymbol{D}^T = [d_1, \cdots, d_K]^T$ and $\bar{\boldsymbol{D}}^T = [\boldsymbol{D}^T, \boldsymbol{F}_1, \cdots, \boldsymbol{F}_L]$. The prior is proper if the integral of the kernel is finite.

All proofs are found in Appendix A and involves rotating $\boldsymbol{\beta}$ such that a flat prior is induced on $p - \text{rank}(\bar{\boldsymbol{D}})$ components of $\boldsymbol{\beta}$. This immediately leads to Theorem 1.

Theorem 1 *For $\lambda > 0$, a prior specified by Definition 1 is proper if and only if $\bar{\boldsymbol{D}}$ is full column rank.*

Unfortunately, Theorem 1 rarely holds in practice as many theoretically motivated choices of $\bar{\boldsymbol{D}}$ are not full rank. Thus, Theorem 2 derives two simple and easily testable conditions for posterior propriety. Proof of sufficiency leverages existing results by Michalak and Morris (2016) on partially flat priors and the orthogonal rotation of $\boldsymbol{\beta}$.

Theorem 2 *Assume a model of the following form:*

- *Likelihood: $L(\boldsymbol{\eta}|\boldsymbol{y}) \equiv f(\boldsymbol{y}|\boldsymbol{\eta})$ where $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$. $\boldsymbol{X} \in \mathbb{R}^{N \times p}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$. Further, assume that the likelihood is log-concave with respect to $\boldsymbol{\eta}$.*

- *Prior: $p(\boldsymbol{\beta}) \propto \exp\left(-\lambda \left[||\boldsymbol{D}\boldsymbol{\beta}||_1 + \sum_{\ell=1}^{L} \sqrt{\boldsymbol{\beta}^T \boldsymbol{F}_\ell \boldsymbol{\beta}}\right]\right)$ as in Definition 1.*

Consider the following conditions;

(a) Augmented Design is Full Rank: $\text{rank}\left(\left[\boldsymbol{X}^T \ \bar{\boldsymbol{D}}^T\right]\right) = p$

(b) Unique MLE of Maximally Sparse Model: $\hat{\boldsymbol{\beta}}_{\mathcal{N}(\bar{\boldsymbol{D}})}$ *exists and is unique.*

$$\hat{\boldsymbol{\beta}}_{\mathcal{N}(\bar{\boldsymbol{D}})} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\max} \ L(\boldsymbol{\eta}|\boldsymbol{y}) \quad \text{s.t.} \quad \boldsymbol{D}\boldsymbol{\beta} = \boldsymbol{0}, \ \boldsymbol{\beta}^T \boldsymbol{F}_\ell \boldsymbol{\beta} = 0 \quad \forall \ell$$

(a) is necessary for posterior propriety; (b) is sufficient for posterior propriety.

Corollary 1 sharpens the results for the most popular likelihoods.

Corollary 1 *If the likelihood is linear or multinomial with a standard link (e.g. logistic or probit), then Conditions (a) and (b) in Theorem 2 are jointly necessary and sufficient for posterior propriety.*

Conditions (a) and (b) have intuitive interpretations. Condition (a) encodes the Bayesian version of the usual "full rank" condition for maximum likelihood estimation to have a unique

solution. In the Bayesian version, a design matrix $(\boldsymbol{X})$ that is not full rank—as is common in high dimensional cases—can still be associated with a proper posterior if either the prior is proper (i.e. $\mathrm{rank}(\bar{\boldsymbol{D}}) = p$) or if, together, they create an augmented design matrix that is full rank. Condition (b) considers the following limiting case; assume all restrictions were binding so the model was the maximally sparse one permitted *given* a particular structure. If the MLE of that model is unique and exists, then the posterior must be proper.

In the case of the agnostic and lattice structures discussed above, Condition (b) involves on checking whether a model with a single treatment indicator (i.e. no heterogeneity) has a unique and finite MLE. In a normal experimental set-up, this is trivially satisfied. In the priority (Type) structure, this involves checking whether a MLE with a six-leveled treatment for project type results in a unique and finite MLE. Again, this is usually satisfied.

Overall, Theorem 2 and Corollary 1 can be easily checked before estimation and thus ensure that researchers can safely use the proposed method.

4.2   Estimating Models with Bayesian Structured Sparsity

Given a proper posterior, estimation in a Bayesian framework typically seeks to sample from the posterior implied by Theorem 2. Theorem 3 facilitates this using data augmentation.

Theorem 3 *Given a structured sparse prior with $\boldsymbol{D}$ and $\{\boldsymbol{F}_\ell\}_{\ell=1}^{L}$ such that the prior is proper, the following joint density preserves a marginal structured sparse prior on $\boldsymbol{\beta}$.*

$$p\left(\boldsymbol{\beta}, \{\tau_k^2\}_{k=1}^{K}, \{\xi_\ell^2\}_{\ell=1}^{L} | \lambda\right) \propto \frac{\exp\left(-\frac{1}{2}\boldsymbol{\beta}^T\left[\sum_{k=1}^{K}\frac{\boldsymbol{d}_k\boldsymbol{d}_k^T}{\tau_k^2} + \sum_{\ell=1}^{L}\frac{\boldsymbol{F}_\ell}{\xi_\ell^2}\right]\boldsymbol{\beta}\right)}{\prod_{k=1}^{K}\frac{\exp(-\lambda^2/2 \cdot \tau_k^2)}{\sqrt{\tau_k^2}} \prod_{\ell=1}^{L}\frac{\exp(-\lambda^2/2 \cdot \xi_\ell^2)}{\sqrt{\xi_\ell^2}}} \tag{4}$$

Appendix C shows that posterior can be sampled using a Gibbs Sampler and a linear likelihood. It further shows how data augmentation can be used to make binomial and multinomial outcomes easily tractable.

14

Additionally, recall that a motivating rationale for Bayesian estimation was the difficulty of finding a penalized MLE—corresponding to the posterior mode. Even if Bayesian estimation is preferred for many analyses, having the ability to find a penalized MLE is useful for prediction tasks involving cross-validation and calibration of the fully Bayesian model as discussed in Appendix B. Appendix C notes how Theorem 3 immediately leads to a stable and tractable estimation method for a penalized MLE using Expectation Maximization (Dempster, Laird, and Rubin 1977) that works for *all* structures and many non-linear likelihoods. This extends a smaller literature that uses an EM approach to estimate sparse models (e.g. Figueiredo 2003; Polson and S. L. Scott 2011; Ratkovic and Tingley 2017).

5   Simulations on Estimating Heterogeneous Effects

I show the use of structured sparsity by first illustrating a claim earlier in the paper that existing methods fare poorly when the underlying heterogeneous effect is represented by *groups* of distinct values. Table 2 outlines a simple simulation environment to test this.

It differs from existing simulations insofar as most units have an effect of either '1' or '-1' although which unit is in which group is unknown. This means that a global or aggregate effect is not representative of many units. This is a plausible case to describe real patterns of heterogeneous effects and thus it is important to understand how existing methods fare in this scenario. I also vary the amount of "grouped" structure by considering a more traditional set-up where many units have a treatment effect of zero.

To begin, I illustrate the results from a handful of simulations ($r = 20; G = 25$)to show the differences between structured sparsity and classical methods. Figure 2 presents the estimated heterogeneous effects from a fixed effects, random effects, and (adaptive) structured sparse model across four replications. Note that the points are *arbitrarily* re-ordered such that the units with a negative effect are on the left and a positive effect are on the right. None of the estimated models know the true groups beforehand. In a model with

Table 2: Simulation Environment

- Parameters: G units; r observations per group. This implies $N = G \times r$ observations.

- Treatment Vector: Assume that $\mathcal{S}$ groups have an effect of '1', $\mathcal{S}$ have an effect of '-1' and $G - 2\mathcal{S}$ have an effect of zero. Without loss of generality, assume the groups are ordered such that the first $\mathcal{S}$ have an effect of '-1' and the last $\mathcal{S}$ have an effect of '1'.

$$\boldsymbol{\tau} = [\underbrace{-1, \cdots, -1}_{g \in \{1,2,\cdots,\mathcal{S}\}}, \underbrace{0, \cdots, 0, \cdots, 0}_{g \in \{\mathcal{S}+1,\cdots,G-\mathcal{S}\}}, \underbrace{1, \cdots, \cdots, 1}_{g \in \{G-\mathcal{S}+1,\cdots,G\}}]$$
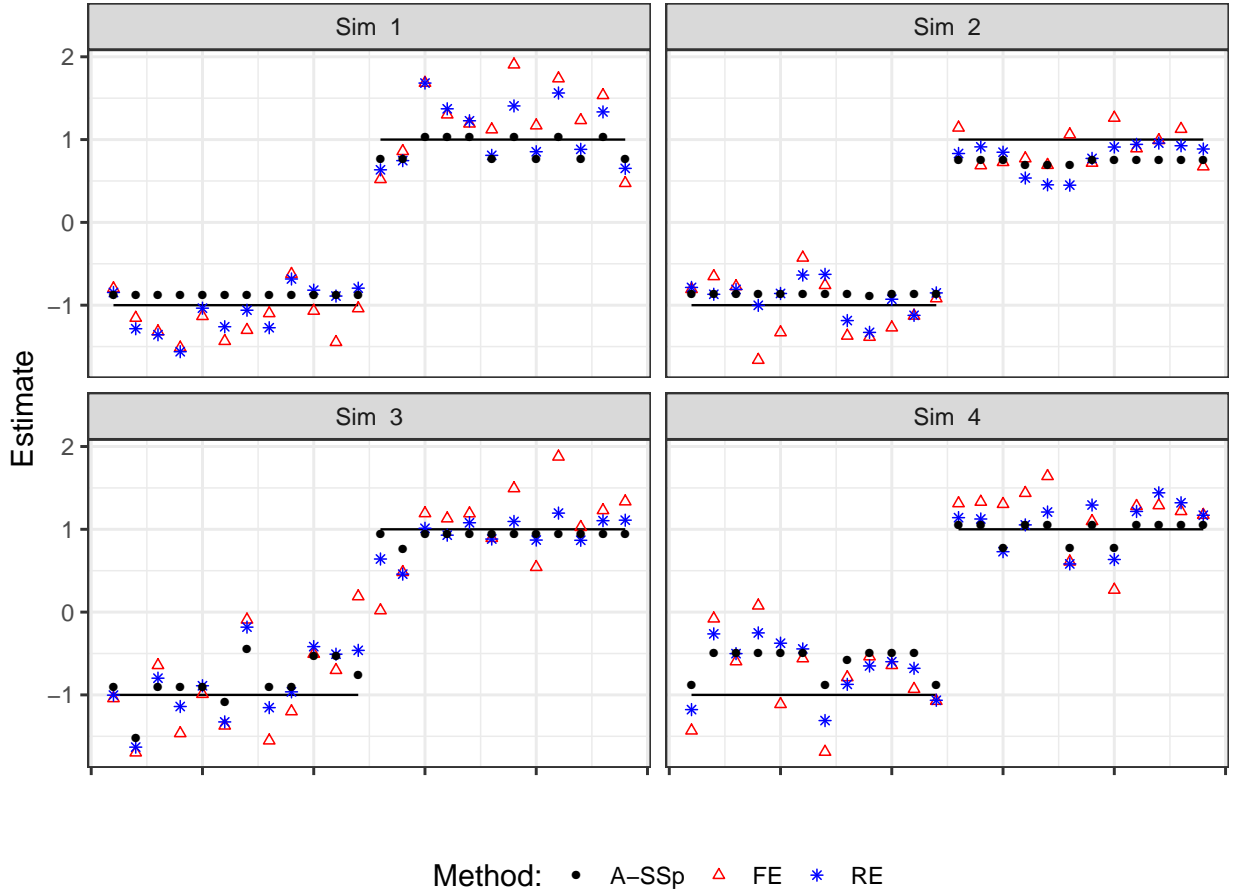
- Outcome:

  - The generative model is: $y_i = x_i + \tau_{g[i]} d_i + \epsilon_i; \quad x_i \sim N(0,1); \quad \epsilon_i \sim N(0,1)$
  - Half of the units in each group are randomly assigned to treatment ($d_i = 1$).

- Models: All models are correctly specified, i.e. include a control for x_i and the possibility of estimating heterogeneous effects for each group g. For example, the LASSO model includes x_i and interactions between d_i and indicator variables for each group. Appendix D outlines the full specification for each model.

perfect performance, all estimates would lie on the corresponding solid line.

The figure provides some intuition as to why stabilization is necessary and structured sparsity out-performs traditional methods: First, note that random effects performs better than fixed effects; fixed effects has around a 30% worse RMSE averaged across the four plots in Figure 2. However, neither approach meaningfully captures a "group" structure, although both usually show clear separation between the units with positive and negative effects. By contrast, structured sparsity clearly out-performs both noticeably and does a quite good job of recovering the unknown-but-underlying groups (fixed effects having a 95% larger RMSE than structured sparsity; 44% for random effects). While the grouping is not perfect, as expected given finite data, it allows for both a more interpretable result (i.e. the units are more often clearly separated) and more accuracy by more effectively pooling information. Its ability to recover the unknown-but-underlying groups shows its power and promise in this situation.

Table 3 examines this more systematically by showing the RMSE of the heterogeneous

Figure 2: Visualizing Selected Simulations

Note: The estimated heterogeneous effect for each unit is shown; the units are arbitrarily to be in order of their true value. The three methods are fixed effects (FE), random effects (RE), and adaptive structured sparsity (A-SSp). The exact specification is described in Appendix D. The true value for each point is shown with a thin black line. For clarity, the one unit with a true zero value is omitted.

effects vs. the truth across eight key methods. It corroborates the benefit of using a regularization method based around groups. Consider first the upper panel where the underlying heterogeneity is indeed grouped (i.e. half of the heterogeneous effects are '1', half are '-1'). For all cases, structured sparsity performs the best. Even using a non-adaptive LASSO, this still outperforms all other methods although the adaptive LASSO weights give considerable performance gains. This confirms that it can effectively exploit group-structured heterogeneity when that well-describes the data generating process.

Of the non-structured sparse methods, a simple hierarchical model with random slopes

17

Table 3: Estimating Heterogeneous Effects for $G = 25$ Units

Mostly Grouped Effects ($\mathcal{S} = \lfloor G/2 \rfloor$)

r	SSp	A-SSp	FE	RE	LASSO	ENet 1	FindIt	BART
10	0.411	**0.401**	0.630	0.427	0.565	0.562	0.443	0.970
	(0.008)	(**0.011**)	(0.009)	(0.007)	(0.007)	(0.006)	(0.006)	(0.001)
20	0.282	**0.229**	0.445	0.313	0.407	0.416	0.315	0.920
	(0.005)	(**0.006**)	(0.006)	(0.005)	(0.005)	(0.005)	(0.005)	(0.002)
50	0.180	**0.115**	0.277	0.202	0.262	0.275	0.196	0.535
	(0.003)	(**0.004**)	(0.004)	(0.003)	(0.004)	(0.004)	(0.003)	(0.007)
100	0.134	**0.078**	0.197	0.145	0.189	0.199	0.143	0.294
	(0.002)	(**0.002**)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)

Mostly Sparse Effects ($\mathcal{S} = \lfloor G/4 \rfloor$)

r	SSp	A-SSp	FE	RE	LASSO	ENet 1	FindIt	BART
10	**0.408**	0.469	0.633	0.414	0.478	0.473	0.438	0.692
	(**0.006**)	(0.007)	(0.010)	(0.007)	(0.008)	(0.008)	(0.007)	(0.002)
20	**0.288**	0.350	0.439	0.296	0.353	0.351	0.304	0.655
	(**0.004**)	(0.007)	(0.006)	(0.004)	(0.005)	(0.006)	(0.005)	(0.002)
50	0.191	**0.176**	0.278	0.203	0.255	0.260	0.198	0.430
	(0.003)	(**0.005**)	(0.004)	(0.003)	(0.005)	(0.004)	(0.003)	(0.005)
100	0.134	**0.098**	0.198	0.143	0.183	0.187	0.136	0.226
	(0.002)	(**0.003**)	(0.003)	(0.002)	(0.004)	(0.004)	(0.002)	(0.004)

Note: The table reports the RMSE for the estimated heterogeneous effects versus the true values averaged across 100 simulations. The standard error is shown in parentheses. r refers to the number of observations per group. The top panel shows the case where most effects are non-zero; the bottom shows mostly zero effects. See Table 2 for details. Appendix D details the methods and software used. The abbreviations stand for, in order, structured sparsity (SSp); an adaptive LASSO version (A-SSp); an interactive model (fixed effects - FE); a hierarchical model with random slopes (RE); LASSO; Elastic Net with $\alpha = 0.5$ (ENet1); FindIt (Imai and Ratkovic 2013); Bayesian Additive Regression Trees (BART).

for treatment by group performs very well followed by traditional sparse methods and FindIt. The poor performance of tree-based methods (BART) suggests that they struggle in settings with categorical variables containing many levels (see Appendix D for more discussion).

Consider the second case (the bottom panel; $\mathcal{S} = \lfloor G/4 \rfloor$) where the heterogeneity is much less grouped, i.e. half of the units have zero effect, one quarter have an effect of '1' and one quarter have an effect of '-1'. First, many traditional methods see improved performance, especially in the case of LASSO and Elastic Net for small r. This makes sense given that their underlying assumptions are more satisfied. Correspondingly, structured

sparsity does worse. The adaptive version sees a noticeably degradation of performance, although it still remains the best performing method for moderately large r. Non-adaptive structured sparsity has stable performance and beats most other methods at all r, although the amount of improvement is certainly more limited.

Taken together, the implications of these simulations is (i) structured sparsity can provide large gains in performance over traditional methods when the underlying heterogeneity is based on groups but (ii) has the ability to flexibly accommodate those scenarios leads to some cost in performance when the traditional methods have the correct assumption (mostly sparse effects). However, since structured sparsity can accommodate the traditional assumption as a special case, it still performs reasonably well. Given that it is not possible in real data to know which assumption is closer to the truth, structured sparsity is most effectively deployed when there is a reasonable suspicion that a grouped analysis might be effective to the question at hand but, unlike with random effects or models based on interactions, the groups cannot be precisely specified *ex ante*.

Appendix D provides additional results; first, it visualizes simulations in terms of percentage change in RMSE vs. adaptive structured sparsity to illustrate relative performance and adds four additional methods (SVM, random forest, ridge regression, another elastic net). Re-interpreting Table 3 in this light confirms the substantively size of the improvement: In the mostly grouped case, random effects and FindIt when $r > 10$ RMSEs that are often 40% larger than adaptive structured sparsity and much worse (e.g. 80-100%) when r is large. In the mostly sparse case, the magnitude of improvements is smaller (e.g. only 20-50% when r is large) and that it is sometimes about 10% worse than existing methods.

Second, I examine results with binary outcomes. Structured sparsity remains the best method in the mostly grouped case (although its magnitude of improvement is smaller). Its performance also degrades somewhat more sharply in the non-grouped case although for large r consistently beats competitor methods by small but distinguishable margins.

Third, I vary the number of units to range from five to one-hundred. Except for the

case of very few units ($G = 5$), structured sparsity continues to perform well. To summarize the eighty simulation environments (e.g. outcome type, r, G, and $\mathcal{S}$ combinations) in brief, structured sparsity and its adaptive analogue are in the top-three methods in 74 and 60 cases, respectively.

6 Uncovering Heterogeneous Effects of Credit-Claiming

As discussed previously, the experimental design in Grimmer, Westwood, and Messing (2014) contains five treatment factors and two variables (party identification and ideology) for respondent-specific heterogeneity. The survey contains only 1,074 observations and thus the ability to look for all possible combinations of treatment and respondent is significantly constrained. The existing literature suggests that the type of project is critically important (Grimmer, Westwood, and Messing 2014; Grimmer, Messing, and Westwood 2017), and thus I generate all pairwise interactions between project type and each other type of treatment as well as the interactions with the respondent characteristics. The formula to generate the design is expressed below.

```
~ Type * (Money + Stage + Sponsor + Co-Sponsor) * (Party + Ideology)
```

This leads to a design with 538 parameters and an intercept.[4] A key part of structured sparsity is using one's theory to decide how these should be possibly grouped together. I consider three possible strategies each reflecting increasingly strong theoretical knowledge. Table 4 provides a brief summary as they are described in detail in Section 2.

The agnostic and lattice structures are generated straightforwardly. As noted above, their limiting case as $\lambda \to \infty$ is a model with only a single treatment indicator (i.e. no effect heterogeneity). The third structure ("priority") uses the prior research on this topic to focus on the role of project type. It connects together all effects that contain the same type of

---

4. This is larger than $p = 280$ used in Grimmer, Messing, and Westwood (2017)'s as structured sparsity does not require the specification of an (arbitrary) baseline category for each treatment that may affect performance of, say, LASSO-based methods.

Table 4: Description and Performance of Structures For Credit-Claiming Experiment

Name	Description	Limiting Case	WAIC	CV
Agnostic	Connect all	Single treatment	1294.7	0.451
Lattice	Connect if they share any attribute	Single treatment	1295.3	0.451
Priority	Connect to project "type"	Single factor	1309.8	0.457

Note: This table briefly describes each of three structures used in the analysis. The final columns (WAIC and CV) are measure of fit discussed in the main text; smaller values indicate better performance.

project (e.g. Roads, Roads-$50k, Roads-$50k-Democrat). For terms without a project type, they are connected based on sharing the same treatment factor (e.g. $50k, $50k-Democrat, $20million). Controls for respondent party and ideology are also included additively. Its limiting case is where only the factor for project type was included alongside the respondent controls. Appendix E provides details on the calibration of λ and posterior diagnostics.[5]

To choose between the structures, one can adopt either a theoretical or data-driven approach. My theoretical prior is to use the lattice structure as that provides some coherence to the groups while also not necessarily prioritizing project type to the exclusion of all others. A data-driven approach compares the fit of the models while noting that they may have different complexities.

Table 4 reports two criteria. First, the Widely Applicable Information Criterion (WAIC; Gelman, Hwang, and Vehtari 2014) is a technique for comparing non-nested Bayesian models and is designed to approximate cross-validation. It can be interpreted like a standard information criterion (AIC or BIC) where smaller indicates a better fit. The evidence from this statistic suggests that the agnostic and lattice structures out-perform the priority structure but otherwise are similar. Second, using the fast EM algorithm and the (approximate) AIC, I conduct 20-fold cross-validation (used again in Section 7) and report the cross-validated root mean squared error. This statistic suggests nearly identical performance between the

---

5. I run the model for 4 chains with 5,000 samples after 5,000 burnin. All Gelman-Rubin diagnostics on $\boldsymbol{\beta}$ are good; there is sometimes poor mixing on λ. I show that fixing λ at an optimal value returns nearly identical results.

agnostic and lattice structures although their individual predictions are not perfectly correlated.

Thus, since theory suggests the lattice structure and the data-driven approach is roughly indifferent, I report results using the lattice structure. As there is always researcher discretion over the choice of structure, it is important to examine sensitivity to this. Appendix E compares the estimated heterogeneous effects for each treatment-respondent combination. The agnostic and lattice estimates are nearly perfectly correlated ($\rho = 0.99$), while the priority structure recovers a somewhat distinct pattern ($\rho = 0.95$).

6.1 Interpreting High-Dimensional Heterogeneity

Interpretation of the estimated heterogeneity is complicated as there are nearly 2,000 treatment/respondent effects that can be predicted from the fitted model. For complex models such as random forests or ensembles, initial exploration often begins by calculating the types of profiles that have the largest effects and looking for commonalities.

Structured sparsity as well as other LASSO-based methods have the useful property of allowing a different initial exploratory step as the model itself returns parameter estimates interpretable (roughly) as in a generalized linear model. It is sensible to run the fast EM model to find the penalized MLE and tune λ as noted earlier. The posterior mode will contain exact groups and it is possible to quickly ascertain which effects appear to be meaningfully distinct. Table 5 does this by reporting the largest parameter estimates. It notes that, as is common for many applications, there is a large group that contains most of the parameters (estimated at 0.0486) with the others being located in mostly singleton clusters. The group-based stabilization occurs by regularizing effects to be close to their *neighbors* versus being pulled towards zero as is traditional.

It clearly illustrates that certain types of interactions appear to be driving the estimated heterogeneity. As noted in prior research (Grimmer, Messing, and Westwood 2017), respondent characteristics (party and ideology) appear very important in moderating the effects

22

Table 5: Largest Parameter Estimates at Posterior Mode

(a) Most Positive | | (b) Most Negative |
| --- | --- | --- | --- |
| Type(PP) * Ideology(Lib) | 0.72 | Type(Gun) | -1.52 |
| Sponsor(Rep) * Party(Rep) | 0.56 | Type(Gun) * Party(Dem) | -1.14 |
| Type(Gun) * Ideology(Cons) | 0.46 | Type(Parks) * Co-Sponsor(Dem) | -0.53 |
| Stage(Secure) * Ideology(Lib) | 0.26 | Type(PP) * Ideology(Cons) | -0.46 |
| Money($50thousand) * Party(Rep) | 0.25 | Type(Gun) * Ideology(Lib) | -0.37 |
| Party(Dem) | 0.24 | Co-Sponsor(Rep) * Ideology(Mod) | -0.33 |
| Type(Roads) | 0.24 | Type(Police) * Sponsor(Rep) * Ideology(Lib) | -0.24 |
| Type(Roads) * Sponsor(Rep) | 0.24 | | |
| Type(Roads) * Co-Sponsor(Dem) | 0.24 | Stage(Request) | -0.14 |
| Type(Fire) | 0.23 | Sponsor(Dem) * Ideology(Cons) | -0.11 |
| | | Sponsor(Dem) * Party(Rep) | -0.09 |

Note: 487 of the 538 parameters are grouped together at 0.0486. The top ten largest and smallest parameter estimates are shown above where " * " indicates an interaction. Planned Parenthood is abbreviated to PP.

of the credit-claiming proposals. Further, it immediately illustrates the substantive point where the *combination* of a co-partisan sponsor and the respondent's party or ideology is highly relevant. I examine each of these in turn.

First, the structured sparse approach resolves an odd feature of the results presented in Grimmer, Messing, and Westwood (2017, p. 428): They examine treatment effect by project type, amount of funding or status of project, and respondent ideology. Close inspection reveals that certain types of treatments have the following effect pattern: Conservatives and liberals have more extreme treatment effects in the *same* direction than moderates. While this could be true in theory, the presented results are implausible.

Specifically, it suggests that when reviewing proposals to claim credit for Planned Parenthood, both conservatives and liberals have a *more positive* response than moderates. A similar problem occurs for proposals to create gun ranges where conservatives and liberals are seen to have a *more negative* response than moderates. Given the polarization of American politics around these issues, neither estimated effect seems likely to generalize to the broader population as the prior expectation would be a (weak) ordering of effects by ideology.

This problem does not appear in their analysis of marginal effect of project type by

party (Grimmer, Messing, and Westwood 2017, p. 427) but arises in both analyses where interactive treatment effects are considered and thus may be a "paradox" resulting from marginalization. This could also be statistical noise, but the ensemble approach used in that paper cannot speak to this point.

Figure 3 explores this using structured sparsity. After calculating the analogous quantity, the point estimates for Planned Parenthood and Gun Ranges are sensible.[6] Across all other types of treatment factors, the posterior medians have a sensible ideological ordering of liberals, moderates, and conservatives for both Planned Parenthood and gun ranges. The interactive treatment effects thus recover the pattern in the marginal effects. Looking across other issues, there also appears to be little evidence of moderates having more extreme treatment effects that is distinguishable from statistical noise.
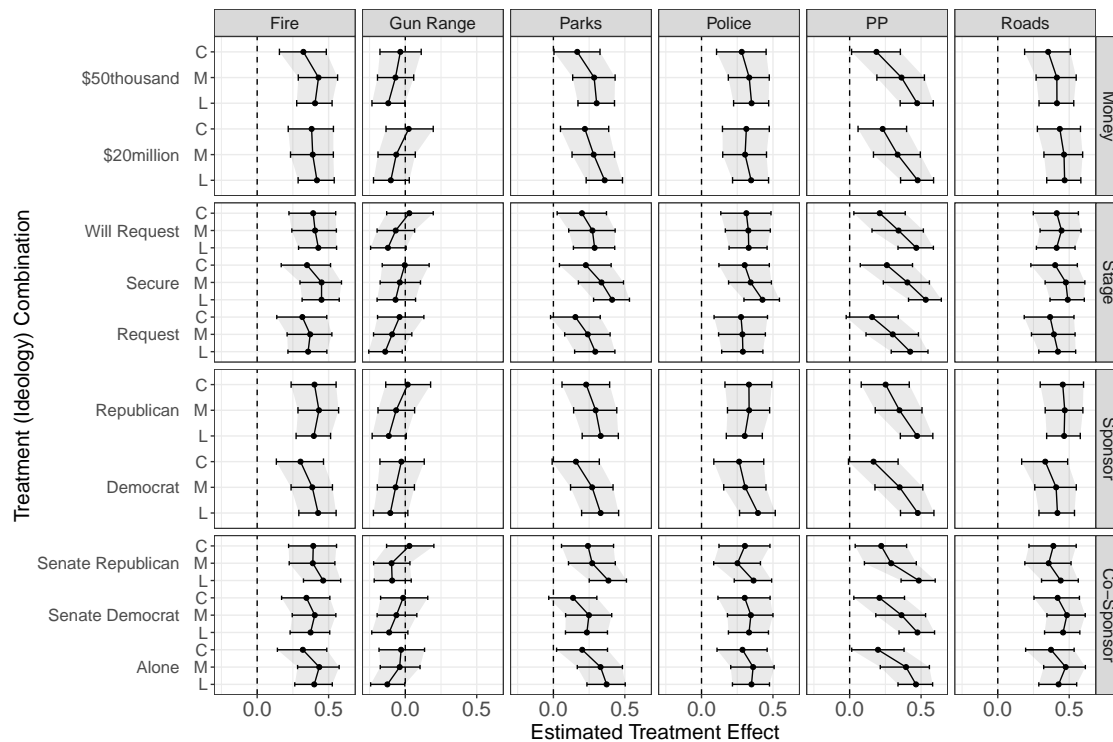
The second substantive finding from structured sparsity concerns the possibility of an interactive effect between the party of the representative proposing the project and the respondent. In their original analysis, Grimmer, Westwood, and Messing (2014, pp. 104-105) examine the role of this effect in a model that does not estimate complex heterogeneity. They find evidence for the role of a larger effect when the respondent has a co-partisan sponsor, although the difference does not always rise to the conventional level of statistical significance. I re-examine this and focus on the differential co-partisan effect by the party and ideology of the respondent as suggested by Table 5. Figure 4 tests this by calculating the conditional average marginal effect of having a Republican sponsor vs. a Democratic sponsor.[7]

It confirms the initial results from analyzing the posterior mode; conservative Republicans show a noticeable preference for proposals from co-partisan representatives (Republicans) over Democratic proposers that is statistically distinguishable from zero. Liberal

---

6. Formally, fix ideology, project type and one other treatment factor at say (liberal, Fire, and $50 thousand). Find all heterogeneous effects with that combination of characteristics and average across them.
7. Formally, fix party and ideology at some combination, say Moderate Independent. Next, calculate the difference in treatment effect between having a Republican and Democratic sponsor for every combination of other treatment factors and that demographic profile. Averaging across those combinations returns the average marginal effect.

Figure 3: Average Treatment Effects by Project Type, Secondary Treatment, and Ideology
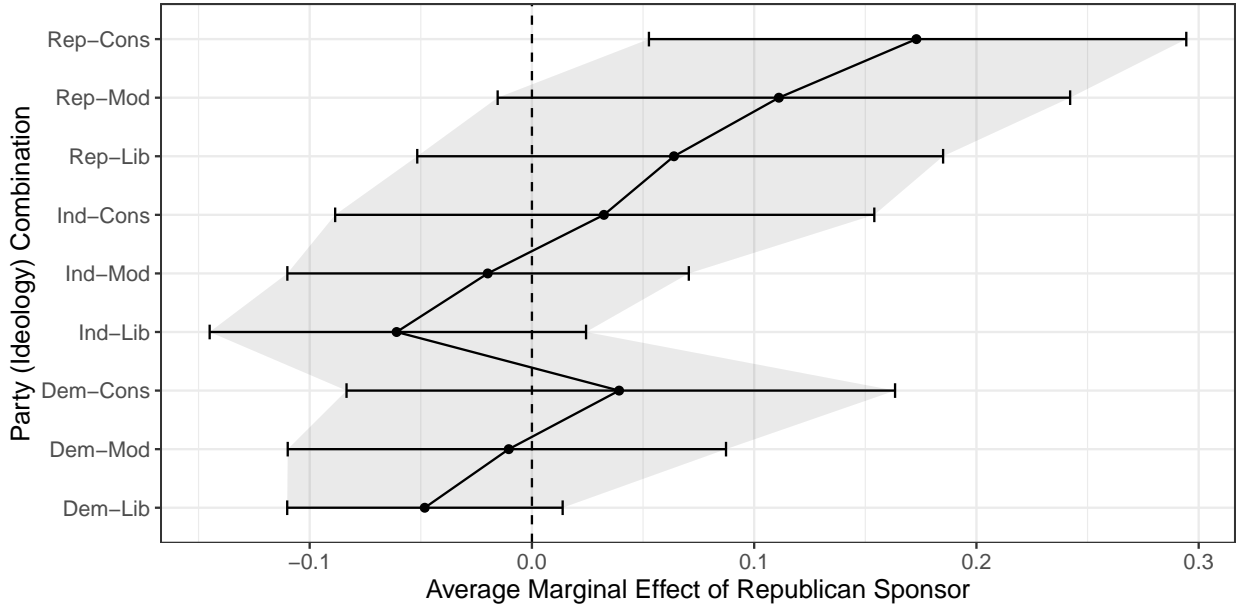

Note: This figure shows the treatment effect for project type (vertical panels) and other treatment factor (horizontal panel) by ideology of the respondent. 'C', 'M', and 'L' denote Conservative, Moderate, and Liberal respondents, respectively. PP stands for Planned Parenthood. The 90% credible interval is reported.

Democrats show the opposite bias (preferring Democratic sponsors), although the credible interval contains zero. This provides an interesting qualification of the results from Grimmer, Westwood, and Messing (2014): There is a statistically distinguishably larger effect for co-partisans evaluating a representative's attempt to claim credit, although it is concentrated in respondents who share the dominant ideological beliefs of their party.

7   Structured Sparsity in an Ensemble

A final test for structured sparsity is how well it performs against other methods for detecting heterogeneous effects on real data. Grimmer, Messing, and Westwood (2017) combine a variety of methods using the idea of "super learning" (Van der Laan, Polley, and Hubbard 2007). This approach fits a variety of methods using K-fold cross-validation and then ex-

Figure 4: Average Marginal Effect of Republican Sponsor by Party and Ideology

Note: This figure shows the difference between having a Republican and Democratic sponsor of the project by party and ideology combinations. The 90% credible interval is reported.

amines how well each method's out-of-sample predictions correspond to the truth. It gives weights to each method that are non-negative and sum to one. Higher weights indicate that the model is used more heavily in the final predictive model and thus is seen to be better at some combination of predicting the outcome and/or being distinct from other methods.

Table 6 shows how structured sparsity fares using a 20-fold split of the data. I replicate the exact collection of models in the ensemble in Grimmer, Messing, and Westwood (2017) and add structured sparse models from Table 4. I also provide a "harder" test by including additional methods (Columns 3-4).[8]

It returns strong support for structured sparsity; in both settings when it is added, two structured sparse models (agnostic and lattice) receive substantial non-zero weight—comprising nearly three-quarters of the total weight in the ensemble. The fact that both models get some weight suggests they are each contributing something distinct and use-

---

8. Note that a random effects model cannot be easily fit here. It would require around *thirty* random effects to capture the proposed interactions (seven for the main effects plus around twenty for the interactions) and thus is prohibitively expensive. Thus, while Section 5 suggests that random effects can perform well, an additional benefit of structured sparsity is its scalability.

Table 6: Ensemble Analysis

Model	Initial Specification (1)	(2)	Structured Sparsity (3)	(4)
LASSO	0.727	0.534	0.246	0.251
ENet1	-	-	-	-
ENet2	-	-	-	-
BayesGLM	-	-	-	-
FindIt	0.273	-	-	-
KRLS	-	-	-	-
SVM	-	-	-	-
Boost		-		-
BART		-		-
RF		0.107		0.013
GLM		0.359		-
SSp (A)			0.281	0.285
SSp (L)			0.473	0.452
SSp (P)			-	-

Note: Appendix D describes the formula and software to fit all models. In addition to models in Table 3, this table includes an Elastic Net with $\alpha = 0.25$ (ENet2), boosting (Boost), Kernel Regularized Least Squares (KRLS); a model with a weakly informative prior (BayesGLM), a support vector machine and a simple logistic regression with each treatment factor and covariate added linearly (GLM). SSp(A), SSp(L), SSp(P) refer to the agnostic, lattice, and priority structures described in Table 4 respectively. The ensemble is constructed using 20-fold cross-validation. $-$ indicates a model was given zero weight.

ful to the ensemble. This again suggests the utility of examining multiple structures. It further suggests that while structured sparsity is not the optimal model in all cases, it is worth including in ensemble analyses as it can cover the possibility of there being grouped heterogeneity in the underlying data.

8 Conclusion

This paper began by noting that, when modelling heterogeneity, it is often necessary to make simplifying assumptions to obtain relatively precise estimates. A vast array of methods provide different sets of assumptions to do this, mostly focused around stabilizing estimates around some global or aggregate effect. The fundamental contribution of this paper is

allowing researchers to use a different assumption: Stabilize effects by creating data-driven groups guided by prior knowledge. Given that the creation of groups is fundamental to how researchers create concepts and make sense of the world, this method is applicable to many research questions across subfields—especially when the proposed model includes many binary or categorical variables.

To do this, existing methods for structured sparsity needed substantial modification to be suitable for social scientific research. The paper developed a Bayesian formulation of structured sparsity that allowed for the quantification of uncertainty in the estimated effects as well as tractable inference for non-linear models with arbitrary penalties. After deriving novel theoretical results about the propriety of the resulting posterior and new inferential techniques, I showed that structured sparsity performed well in two contexts.

First, I showed that using a simple simulation where the underlying heterogeneous effects fell mostly into two distinct groups, state-of-the-art methods struggled to accurately estimate the heterogeneous effects. Their approach of shrinking aggressively towards a global effect failed to do well insofar as that global effect did not represent many groups. By contrast, structured sparsity—as well as traditional methods such as hierarchical models—performed well. The key benefit of structured sparsity over traditional methods such as random effects is that it can *estimate* the group memberships, while being guided by theory, rather than requiring them to be exactly specified *ex ante*.

Second, I applied Bayesian structured sparsity to a recent experiment on credit-claiming and drew out two new substantive implications. First, an unusual published result of moderates being more ideologically extreme for certain combinations of treatments did not occur when using structured sparsity. For salient policies such as building gun ranges or Planned Parenthood clinics, structure sparsity estimated marginal as well as interactive treatment effects that agreed with the dominant understanding of American politics.

Guided by the interpretable output from structured sparsity, I examined the claim that respondents react more favorably to policies proposed by a co-partisan representative (Grim-

mer, Westwood, and Messing 2014). I also found qualified evidence for this but noted that the effect appears concentrated in conservative Republicans and liberal Democrats, although only the former shows effects that were statistically distinguishable from zero. Finally, I compared structured sparsity against a wide array of state-of-the-art methods in an ensemble, I found that structured sparsity had the best cross-validated error and received significant non-zero weight in an ensemble of methods stacked together.

Overall, the paper demonstrates a dual role to structured sparsity; first, if a single model is desired and one believes that *groups* may represent a reasonable way of representing the underlying heterogeneity, it is a flexible general purpose tool for uncovering easily interpretable heterogeneity while quantifying uncertainty. Second, if one wishes to rely on a collection of models, its inclusion into an ensemble can account for the possibility of grouped heterogeneity and thus improve the performance of this power technique.

References

Arnold, Taylor B., and Ryan J. Tibshirani. 2016. "Efficient Implementations of the Generalized Lasso Dual Path Algorithm." *Journal of Computational and Graphical Statistics* 25 (1): 1–27.

Bach, Francis, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. 2012. "Structured Sparsity through Convex Optimization." *Statistical Science* 27 (4): 450–468.

Betancourt, Brenda, Abel Rodríguez, and Naomi Boyd. 2017. "Bayesian Fused Lasso Regression for Dynamic Binary Networks." *Journal of Computational and Graphical Statistics* 26 (4): 840–850.

Bondell, Howard D., and Brian J. Reich. 2009. "Simultaneous Factor Selection and Collapsing Levels in ANOVA." *Biometrics* 65 (1): 169–177.

Bonhomme, Stéphane, and Elena Manresa. 2015. "Grouped Patterns of Heterogeneity in Panel Data." *Econometrica* 83 (3): 1147–1184.

Chen, Xi, Qihang Lin, Seyoung Kim, Jaime G. Carbonell, and Eric P. Xing. 2012. "Smoothing Proximal Gradient Method for General Structured Sparse Regression." *The Annals of Applied Statistics* 6 (2): 719–752.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *The Econometrics Journal* 21 (1): C1–C68.

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)* 39 (1): 1–38.

Fan, Jianqing, and Runze Li. 2001. "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties." *Journal of the American Statistical Association* 96 (456): 1348–1360.

Faulkner, James. R, and Vladimir N. Minin. 2018. "Locally Adaptive Smoothing with Markov Random Fields and Shrinkage Priors." *Bayesian Analysis* 13 (1): 225–252.

Figueiredo, Mário A.T. 2003. "Adaptive Sparseness for Supervised Learning." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25 (9): 1150–1159.

Gaines, Brian R., Juhyun Kim, and Hua Zhou. 2018. "Algorithms for Fitting the Constrained Lasso." *Journal of Computational and Graphical Statistics* 27 (4): 861–871.

Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press.

Gelman, Andrew, Jessica Hwang, and Aki Vehtari. 2014. "Understanding Predictive Information Criteria for Bayesian Models." *Statistics and Computing* 24 (6): 997–1016.

Gertheiss, Jan, and Gerhard Tutz. 2010. "Sparse Modeling of Categorial Explanatory Variables." *The Annals of Applied Statistics* 4 (4): 2150–2180.

Grimmer, Justin, Solomon Messing, and Sean J. Westwood. 2017. "Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods." *Political Analysis* 25 (4): 413–434.

Grimmer, Justin, Sean J. Westwood, and Solomon Messing. 2014. *The Impression of Influence: Legislator Communication, Representation, and Democratic Accountability.* Princeton, NJ: Princeton University Press.

Hill, Jennifer. 2011. "Bayesian Nonparametric Modeling for Causal Inference." *Journal of Computational and Graphical Statistics* 20 (1): 217–240.

Hobert, James P., and George Casella. 1996. "The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models." *Journal of the American Statistical Association* 91 (436): 1461–1473.

Huang, Junzhou, Tong Zhang, and Dimitris Metaxas. 2011. "Learning with Structured Sparsity." *Journal of Machine Learning Research* 12:3371–3412.

Imai, Kosuke, and Marc Ratkovic. 2013. "Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation." *The Annals of Applied Statistics* 7 (1): 443–470.

Imai, Kosuke, and Aaron Strauss. 2011. "Estimation of Heterogeneous Treatment Effects from Randomized Experiments, with Application to the Optimal Planning of the Get-Out-the-Vote Campaign." *Political Analysis* 19 (1): 1–19.

Jacob, Laurent, Guillaume Obozinski, and Jean-Philippe Vert. 2009. "Group Lasso with Overlap and Graph Lasso." In *International Conference on Machine Learning.* https://www.icml.cc/Conferences/2009/papers/471.pdf.

Kyung, Minjung, Jeff Gill, Malay Ghosh, and George Casella. 2010. "Penalized Regression, Standard Errors, and Bayesian Lassos." *Bayesian Analysis* 5 (2): 369–412.

Leeb, Hannes, and Benedikt M. Pötscher. 2005. "Model Selection and Inference: Facts and Fiction." *Econometric Theory* 21 (1): 21–59.

Ma, Shujie, and Jian Huang. 2017. "A Concave Pairwise Fusion Approach to Subgroup Analysis." *Journal of the American Statistical Association* 112 (517): 410–423.

Mayhew, David R. 1974. *Congress: The Electoral Connection.* New Haven: Yale University Press.

Michalak, Sarah E., and Carl N. Morris. 2016. "Posterior Propriety for Hierarchical Models with Log-Likelihoods That Have Norm Bounds." *Bayesian Analysis* 11 (2): 545–571.

Park, Trevor, and George Casella. 2008. "The Bayesian Lasso." *Journal of the American Statistical Association* 103 (482): 681–686.

Pauger, Daniela, and Helga Wagner. 2019. "Bayesian Effect Fusion for Categorical Predictors." *Bayesian Analysis* 14 (2): 341–369.

Polson, Nicholas G., and James G. Scott. 2011. "Shrink Locally, Act Globally: Sparse Bayesian Regularization and Prediction." In *Bayesian Statistics 9,* edited by José M. Bernardo et al.

Polson, Nicholas G., and Steve L. Scott. 2011. "Data Augmentation for Support Vector Machines." *Bayesian Analysis* 6 (1): 1–24.

Ratkovic, Marc, and Dustin Tingley. 2017. "Sparse Estimation and Uncertainty with Application to Subgroup Analysis." *Political Analysis* 25 (1): 1–40.

Shahn, Zach, and David Madigan. 2017. "Latent Class Mixture Models of Treatment Effect Heterogeneity." *Bayesian Analysis* 12 (3): 831–854.

Shen, Juan, and Xuming He. 2015. "Inference for Subgroup Analysis With a Structured Logistic-Normal Mixture Model." *Journal of the American Statistical Association* 110 (509): 303–312.

Shiraito, Yuki. 2016. "Uncovering Heterogeneous Treatment Effects." Accessed on 21 Aug 2019. https://shiraito.github.io/research/files/jmp.pdf.

Tansey, Wesley, Alex Athey, Alex Reinhart, and James G. Scott. 2017. "Multiscale Spatial Density Smoothing: An Application to Large-Scale Radiological Survey and Anomaly Detection." *Journal of the American Statistical Association* 112 (519): 1047–1063.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (1): 267–288.

Tibshirani, Ryan J., and Jonathan Taylor. 2011. "The Solution Path of the Generalized Lasso." *The Annals of Statistics* 39 (3): 1335–1371.

Tseng, Paul. 2001. "Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization." *Journal of Optimization Theory and Applications* 109 (3): 475–494.

Van der Laan, Mark J., Eric C. Polley, and Alan E. Hubbard. 2007. "Super Learner." *Statistical Applications in Genetics and Molecular Biology* 6 (1).

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association* 113 (523): 1228–1242.

Yuan, Ming, and Yi Lin. 2006. "Model Selection and Estimation in Regression with Grouped Variables." *Journal of the Royal Statistical Society. Series B (Methodological)* 68 (1): 49–67.

Zhu, Yunzhang. 2017. "An Augmented ADMM Algorithm With Application to the Generalized Lasso Problem." *Journal of Computational and Graphical Statistics* 26 (1): 195–204.