Modelling Heterogeneity Using Bayesian Structured Sparsity

Max Goplerud*

September 7, 2019

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.

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

*Draft. Please see `http://mgoplerud.com/papers/Goplerud_Sparsity.pdf` for the latest version. I thank participants at PolMeth 2019, EPSA 2019, New Faces in Political Methodology XI, MPSA 2019, Princeton's Quantitative Social Science Colloquium, Asian PolMeth 2019, APSA 2018, PolMeth 2018, and Text-As-Data 2018 for helpful comments. I thank the following people for comments on earlier versions of the paper: Naoki Egami, Shusei Eshima, June Hwang, Kosuke Imai, Gary King, Dean Knox, Shiro Kuriwaki, Naijia Liu, Ian Lundberg, Walter Mebane, Sun Young Park, Casey Petroff, Tyler Simko, Diana Stanescu, Brandon Stewart, Marc Ratkovic, Dustin Tingley, Jeremy Wachter, and Soichiro Yamauchi provided helpful feedback on earlier drafts of this paper. I thank the Alexander and Diviya Magaro Peer Pre-Review Program at Harvard's Institute for Quantitative Social Science for providing helpful feedback from an anonymous reviewer. David Stadelmann, Marco Portmann, and Reiner Eichenberger kindly shared the underlying data on Swiss MPs. Prior versions of this paper were known as "Modelling Heterogeneity Using Complex Sparsity". All remaining errors are my own.

1 Introduction

Data analysis in social science faces a fundamental trade-off. Whether the researcher focuses on studying voters, elected representatives, organizations, or something else entirely, there is a huge amount of richness and complexity in each observation. Yet, statistical models abstract away from this by making assumptions about the comparability of observations, the constant effects of some variables, functional form assumptions, or other similar ideas. 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 assumption 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. Straightforwardly 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 control covariates included in the analysis (e.g. by interacting indicator variables with the key explanatory variable). Unfortunately, for most datasets, this approach typically leads to unacceptably noisy estimates of the sub-group effects because there are often very few, and perhaps only one, observation with the same covariates. Simplifying the problem by assuming some stability in the underlying heterogeneity versus this "assumption free" approach is a large and active area of research (e.g. "estimating heterogeneous treatment effects" in an experimental context; see, e.g., Imai and Strauss 2011; Athey and Imbens 2016; 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 classical methods such as hierarchical models (e.g. "partial pooling"; Gelman and Hill 2006), sparse models

(e.g. the LASSO; Tibshirani 1996), as well as—more implicitly—in methods based on regression trees (e.g. Hill 2011; Green and Kern 2012; Wager and Athey 2018). A key goal of this paper is to suggest, however, that such an assumption about how to simplify underlying heterogeneity, while very useful in many circumstances, does not reflect a common alternative way of creating understanding. Specifically, the methods in this paper leverage the intuition that humans often understand complex phenomena by classifying observations into a small number of *groups* of distinct observations. Such group creation is ubiquitous in political science research and appears whenever scholars create typologies, categorical variables, etc. A group-based explanation is also a natural way of explaining a complex phenomenon to others and thus methods based on this approach have the benefit of being easily interpretable.

The project of this paper is to provide a quantitative framework for social science research to formalize that intuition. It does this by relying on a method often known as "structured sparsity" to create groups using a combination of prior knowledge and data (e.g. Huang, Zhang, and Metaxas 2011; Bach et al. 2012; Chen et al. 2012).[1] Broadly, it works in the following way: it starts from an agnostic scenario where there are many parameters representing the effect in each small unit (e.g. from interacting indicator variables with a key explanatory variable). 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—based on their prior knowledge of the substantive question at hand. Then, based on the data, the model determines whether two units should be fused together (i.e. given the *same* effect) if (a) their estimated effects are close and (b) prior knowledge permits their combination. From this process, groups emerge from units being given the same estimated 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 scien-

[1]Many papers in this literature use somewhat different terms for similar models (e.g. generalized LASSO; Tibshirani and Taylor 2011; concave fusion Ma and Huang 2017). Appendix A provides a broad overview. Specifically, in this paper, I used structured sparsity to refer to a generalized LASSO (Tibshirani and Taylor 2011) or other model that imposes sparsity on a linear combination of coefficients. The other part of structured sparsity in some uses (e.g. Bach et al. 2012) is known as a "group penalty" (Yuan and Lin 2006) and can be added to the model without difficulty.

tific research. Existing implementations focus on inducing structured sparsity by penalizing the likelihood; as the penalty is non-differentiable, these approaches do not provide a way to quantify uncertainty in the estimated groups or easily propagate it through to the other parameters in the model.[2] This violates one of the key quantitative principles of social science insofar as, for inferential research, accounting for uncertainty is fundamental. Moreover, existing inferential techniques are typically designed for the linear model or particular penalty structures and thus are not suitable for estimating the common types of non-linear models encountered in political science research.

I address these concerns by creating a Bayesian formulation of structured sparsity by allowing for a generalized LASSO with *any* penalty structure to be translated into a prior and remain tractable. This extends existing work that uses particular implementations of structured sparsity (e.g. Kyung et al. 2010; Roualdes 2015; Betancourt, Rodríguez, and Boyd 2017; Tansey et al. 2017; Faulkner and Minin 2018) by focusing on theoretical properties and inference in a general regression context. To do this, I provide novel theoretical results on the propriety of a structured sparse prior and the resulting posterior—as the prior is often improper by design. As the posterior mode is of special interest because it corresponds to the traditional structured sparse estimate, I provide a new EM algorithm to find this posterior mode for linear and non-linear models. This compliments existing methods for estimating the generalized LASSO that are typically focused on linear models and/or specific choices of penalty (e.g. Arnold and Tibshirani 2016; Zhu 2017; Chen et al. 2012; Tansey et al. 2017) and provides speed gains versus existing software (e.g. Zhu 2017).

I demonstrate the usefulness of Bayesian structured sparsity in two ways. First, I use simulations to justify the initial claim in this paper that existing methods for estimating heterogeneous effects perform poorly when the underlying heterogeneity is based around *groups*. Using a simple example where the global effect is not representative of many sub-group effects, I show that many state-of-the-art methods (e.g. BART, LASSO, and others) perform worse than more conventional methods (e.g. 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.

Second, I turn to an empirical application by re-examining a recent paper (Giger and Klüver 2016). As is common, the authors estimate a single coefficient to represent the effect of their key

[2]Kyung et al. (2010) show in the standard sparse case, the bootstrap also does not provide correct quantification of uncertainty; it is likely that similar concerns apply to structured sparsity and thus a "simple" fix is not available.

explanatory variable; I show that exploring heterogeneous by sub-groups (party and type of vote) shows noticeably different results than the original analysis as most of the effect is driven by MPs from major parties.

2 Estimating Heterogeneous Effects

Estimating whether the effect of some key explanatory variable differs by sub-group is a key part of hypothesis testing and data exploration in political science; I focus on re-analyzing a recent paper by Giger and Klüver (2016) to illustrate this point. Their analysis tackles an important substantive question in legislative politics, uses standard methods for observational data, and has already been shown to be a fruitful dataset for exploring heterogeneous effects.[3] Substantively, the authors wish to understand the conditions under which elected representatives vote against their constituents' preferences. Giger and Klüver (2016) use a unique dataset from Switzerland that contains preferences of representatives and voters on identical questions (Stadelmann, Portmann, and Eichenberger 2013) and focus on the role of interest group attachments leading to greater defection. As they observe the votes of all MPs across a number of different referendums (116 from 1995 to 2009; 20260 votes in total), they leverage variation in the number of interest group affiliations across MPs and over time to see whether this corresponds to more or less defection from constituents. Interest group affiliations are measured using an official register where MPs must register the name of all groups of which they are formally affiliated (e.g. having a membership or official function; Giger and Klüver 2016, p. 195). The register is updated yearly and thus there is some within-MP variation in the number of affiliations. Giger and Klüver (2016) examine the register and classify groups into two types: "economic" groups ("sectional" groups in their original parlance; e.g. farmers, chemical associations; p. 193) and "cause" groups (e.g. NGOs). Their key theoretical innovation is to argue that these types of groups should have different effects on MP's defections from their constituents' preferences. They argue that economic groups represent particularistic interests and thus provide incentives for MPs to defect from their constituents (p. 193). On the other hand, cause groups should "strengthen the congruence between legislators and

[3] A number of other papers have used the same dataset to analyze other questions such as district magnitude, margin of victory and others (e.g. Portmann, Stadelmann, and Eichenberger 2012; Carey and Hix 2012; Barceló 2017).

their voters, as they typically support the same policy goals as the majority of citizens" (p. 193). Thus affiliations to cause groups should lead to fewer defections from constituent preferences.

The authors rely on a hierarchical logistic regression (Gelman and Hill 2006) to estimate the key parameters in their model.[4] Equation 1 outlines their original specification. $y_{i,r}$ is a binary variable where '1' represents MP i defecting from their constituents on referendum r. $\boldsymbol{x}_{i,r}$ is a vector of controls described in Appendix F including referendum type, salience of vote, time until next election, and others. The key explanatory variables are the number of interest group attachments for economic and cause groups ($\text{econ}_{i,r}$ and $\text{cause}_{i,r}$, respectively). They also include random effects for the party of the MP (α_p) and canton (α_c).

$$
\begin{aligned}
y_{i,r} &\sim \text{Bernoulli}\left(\frac{\exp(\psi_{i,r})}{1+\exp(\psi_{i,r})}\right) \\
\psi_{i,r} &= \boldsymbol{x}_{i,r}^T\boldsymbol{\beta} + \text{econ}_{i,r} \times \tau_{\text{econ}} + \text{cause}_{i,r} \times \tau_{\text{cause}} + \alpha_{p[i,r]} + \alpha_{c[i,r]} \\
\alpha_p &\sim N(0, \sigma_p^2); \quad \alpha_c \sim N(0, \sigma_c^2)
\end{aligned}
\tag{1}
$$

They interpret the estimated coefficients on the economic and cause group variables ($\hat{\tau}_{\text{econ}}, \hat{\tau}_{\text{cause}}$) and their statistical significance as evidence of an effect of interest group attachments. A key question, however, is whether those two coefficients accurately reflect the underlying effect. For example, one might wonder whether a single coefficient on, say, economic interests masks the fact that only certain types of MPs or votes have a positive and significant effect. Alternatively, it could be the case that some groups see a negative effect and others see a positive effect and thus the weighted average reported in a single number may be misleading. This process of uncovering heterogeneity inside of a single regression coefficient is an important and common goal in political science. Sometimes a specific hypothesis requires testing an interaction with the key variable (e.g. Barceló 2017's re-analysis of Giger and Klüver 2016). Other times, one wishes to uncover heterogeneity more flexibly; I focus on this case and explore whether the effect of interest group attachments differs by the party of the MP, the type of vote, or their interaction. Both are substantively important dimensions of heterogeneity in legislative politics that remain unexplored by existing research on this dataset.

[4] I follow the authors in, implicitly, assuming the relevant assumptions (e.g. selection on observables) hold for the results to be interpreted causally.

Mathematically, both objectives start in the same way by creating interactions with the key variable of interest and the units for which heterogeneous effects are desired. Let $\boldsymbol{z}_{i,r}$ represent a collection of indicator variables (party-and-referendum type combination) for each observation (i, r) in the dataset. For each observation, exactly one element of $\boldsymbol{z}_{i,r}$ equals '1'; the rest equal zero. The model for heterogeneous effects can be created by slightly adapting the systematic component $(\psi_{i,r})$ in the original model. In Equation 2, τ_{econ} represents the global or aggregate effect and $\boldsymbol{\delta}$ represents a vector of deviations from the global effect. The combination of the two for some group represents the estimated heterogeneous effect.

$$\psi_{i,r} = \boldsymbol{x}_i^T \boldsymbol{\beta} + \text{econ}_{i,r} \left[\tau_{\text{econ}} + \boldsymbol{\delta}_{\text{econ}}^T \boldsymbol{z}_{i,r}\right] + \text{cause}_{i,r} \left[\tau_{\text{cause}} + \boldsymbol{\delta}_{\text{cause}}^T \boldsymbol{z}_{i,r}\right] + \alpha_{p[i]} + \alpha_{c[i]} \tag{2}$$

Unfortunately, estimating this model without further assumptions runs into severe difficulty. Even setting aside the fact the maximum likelihood estimate may not be defined because of separation, there are often only a small number of observations for each party-and-referendum type (i.e. each element of $\boldsymbol{z}_{i,r}$). This means that any estimated effects are likely to be very noisy and thus creates difficulties in drawing reliable inferences.

To tackle this problem, a huge and very active literature focuses on developing methods for assuming some stability or structure in the underlying heterogeneity to obtain more precise estimates of the heterogeneous effects. Most of the literature is focused on the context of a randomized experiment and thus discusses estimating heterogeneous treatment effects (see Imai and Strauss 2011; Grimmer, Messing, and Westwood 2017; Athey and Imbens 2016 for reviews from different perspectives), but similar strategies apply to observational data. The diverse and innovative methods in that literature rely fundamentally on the same idea: Given that estimating the most agnostic and flexible model (i.e. an interactive approach such as the one specified in Equation 2) is impossible or leads to unacceptably noisy estimates for the estimated sub-group effects, some simplifying assumptions must be made. The cost of the assumptions is bias, i.e. incorrect estimates of the sub-group effects, at the benefit of gaining considerable efficiency and limiting the risk of incorrectly concluding that some unit has an effect merely due to limited data and random variation. Exactly *how* that trade-off is made depends on the particular method under consideration. A key point of this paper is that, for doing inference, that choice should be substantive and based on the

researcher's belief as to the structure of the underlying heterogeneity.[5]

Most existing methods work on the following simplification; the majority of elements of $\boldsymbol{\delta}_{\text{econ}}$ should be stabilized by pulling them towards zero. This presumes a model where a single global effect can well-describe most observations (i.e. no heterogeneity) and, for many methods, deviations from that global effect are likely small for most units. Exactly how that stabilization is done depends on the model, but this idea characterizes many popular models such as hierarchical random effects models (Gelman and Hill 2006), sparse models (e.g. LASSO, FindIt; Tian et al. 2014; Imai and Ratkovic 2013), and methods based on regression trees (e.g. BART, causal trees or forests; Hill 2011; Athey and Imbens 2016; Wager and Athey 2018).[6]

However, it is sometimes the case that simplifying heterogeneity in a *different* way may be more appropriate. Consider the case where there were two distinct groups—one with positive effects and one with negative effects. In that case, shrinking towards a global effect may be unhelpful as the global effect poorly describes most units. Similar concerns apply if there were multiple groups of observations with distinct effects. Note, however, that if it is true that most units have no heterogeneity (i.e. most elements of $\boldsymbol{\delta}_{\text{econ}}$ are zero), this can be captured by creating one large group containing most observations. Overall, an approach based on groups is naturally designed to address these alternative types of heterogeneity that are—as I demonstrate in Section 5—not well captured by existing approaches. The core premise of this paper is that creating groups of units with the same effect will provide an effective and interpretable way to simplify the agnostic model and estimate heterogeneous effects.

There are many ways to induce group-based heterogeneous effects. They can be roughly divided into two types; clustering-based approaches and agglomerative approaches. Clustering-based approaches typically work in the following way; the research specifies some number of groups in advance and then the model allocates observations into groups. Finite mixture models (e.g. Shen and He 2015; Shahn and Madigan 2017), "group fixed effects" based on k-means (Bonhomme and

[5] A different strategy would be to use an ensemble of methods (e.g. Grimmer, Messing, and Westwood 2017). From that perspective, this paper argues that adding structured sparsity to an ensemble will help insofar as there are patterns of heterogeneity where it outperforms other approaches. Section 5 provides detailed simulations on this point.

[6] The claim about trees needs more elaboration; note that if the tree is placed only on $\boldsymbol{\delta}_{\text{econ}}$—corresponding to a one-hot set of covariates $\boldsymbol{z}_i$, it will select some number of groups to get a different effect. Given that most tree-based methods limit the depth of the tree to prevent overfitting, this means that only a small number of groups will be given a heterogeneous effect versus a global baseline.

Manresa 2015) and the classifier LASSO (Su, Shi, and Phillips 2016) are all examples of this type of approach. The major downside of these methods is two-fold; first, and most importantly, they cannot easily include theory or prior knowledge about which groups are permissible. For example, in the case above, imagine that two units should only be put together if they share *either* party *or* referendum type. In a different setting, one might wish to create spatially contiguous groups if the units corresponded to, say, counties. The approaches above cannot incorporate such constraints without additional modification. Second, they also assume the number of groups in advance! While number of groups can be varied and model fit checked, it means that for any given run of the model, there cannot be any uncertainty on the number of groups.[7] As that is something the researcher might wish to discover—and likely has little prior belief about, this is undesirable.

An agglomerative approach to creating groups addresses these problems. It starts from a bottom-up approach where the model decides whether to give individual units the same effect (e.g. fusing together two elements of $\boldsymbol{\delta}_{\text{econ}}$). From these comparisons, some number of groups emerge by looking at which units are given identical effects. That number is not specified in advance and, in the Bayesian formulation outlined below, can be examined probabilistically. This bottom-up structure also has the advantage of allowing easy incorporation of prior knowledge. When deciding which groups to form, the agglomerative approach looks at all *permissible* pairwise comparisons and decides whether those units should be fused together. If two units should not be directly put together, that pairing can be removed from consideration by the researcher in advance. This paper relies on a specific approach to creating groups—denoted as "structured sparsity" in the spirit of existing methods (e.g. Huang, Zhang, and Metaxas 2011; Bach et al. 2012; Chen et al. 2012)—using the generalized LASSO (Tibshirani and Taylor 2011) or some other continuous penalty (e.g. Ma and Huang 2017).[8] I choose this approach for two reasons; first, other methods in this family are typically computationally demanding (e.g. spike-and-slab priors; Pauger and Wagner 2018). By contrast, structured sparsity via the generalized LASSO has the advantage of being a tractable method for large datasets while maintaining the agglomerative approach to group creation.

[7] Mixture models based on Dirichlet process priors, e.g. Shiraito 2016, are an exception to the second concern insofar as each posterior draw could have a distinct number of groups. Similarly, careful parameterization of the finite mixture model can allow for many groups to be empty and thus effectively allow the number of groups to vary. Neither can easily incorporate prior belief on which units should or should not be allowed to be grouped together, however.

[8] Models in this vein often have slightly different names or specifications; Appendix A outlines the relationship between this paper's use of "structured sparsity" and other related models.

3 Modelling Heterogeneity Using Structured Sparsity

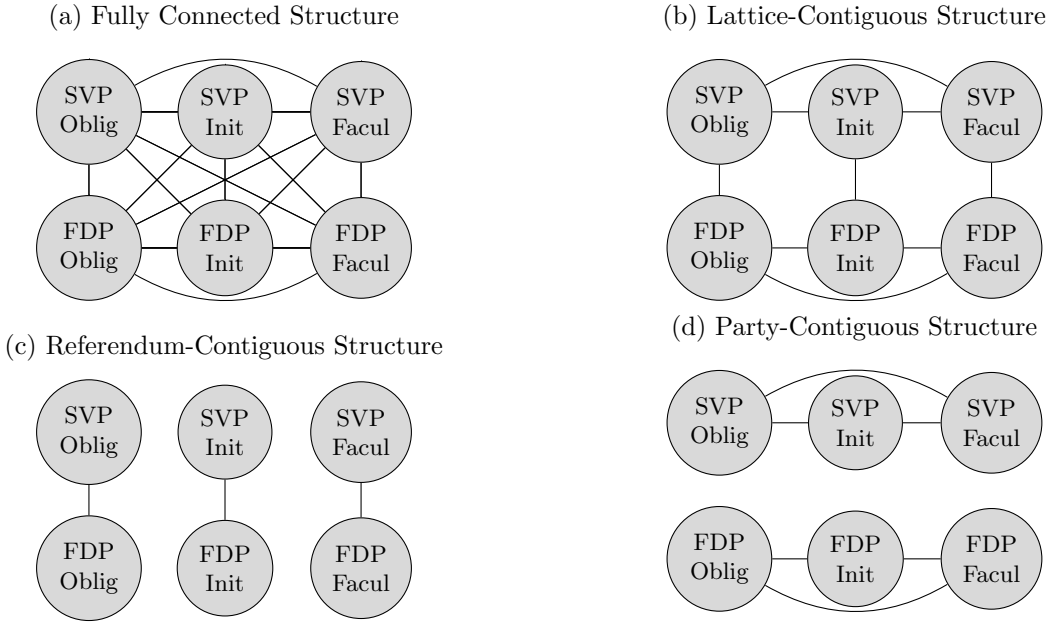
To outline structured sparsity formally, I focus on a reduced slice of the Swiss data with two parties (FDP; Free Democratic Party and SVP; Swiss People's Party) and the three types of referendum found in the data (obligatory [Oblig], facultative [Facul], initiative [Init]). The FDP is a economically liberal party and the SVP is a populist right-wing party; both are represented in the "magic formula" of parties that govern Switzerland (see Kriesi and Trechsel 2008 for an overview of Swiss politics). The three types of referendum are explained in detail in Section 6, but correspond to different ways that the referendum are proposed and the role of the legislature in the process (Giger and Klüver 2016).

The first choice for any research using structured sparsity is how to integrate prior knowledge. This can be visualized most easily by creating a network that connects the units for which heterogeneous effects are being estimated. The idea is to connect units that might be *directly* grouped together. Figure 1 shows four possible structures for this application. First, there is a fully connected structure (1a); this is the most agnostic case. Any groups are possible as any two units could be put together. While it is the most flexible, it does not include any prior knowledge; it therefore may be less efficient 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* party *or* referendum type. Two units might still be in the same group if they are connected indirectly; for example, FDP-Oblig is connected to FDP-Init is connected to SVP-Init. The structure ensures that no group can contain FDP-Oblig and SVP-Init 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 bottom two panels: Referendum and Party. These allow only for groups *within* one dimension of the heterogeneity, e.g. that groups can only be formed by fusing together units with the same party (e.g. FDP-Init and FDP-Oblig). 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 Fully Connected or Lattice structures, then the model becomes one with a single effect—a model with no heterogeneous effects. For the Referendum and Party structures, by contrast, the limiting case is a model with one effect for each referendum type or one for each party, respectively.

Figure 1: Different Types of Structured Sparsity



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 top left panel shows a fully connected structure (all groups are possible); the top right shows a "lattice" structure where two nodes are connected if and only if they share either party or referendum type. The bottom two panels show structures where nodes are connected if they share referendum type or party, respectively.

Estimation using structured sparsity proceeds by examining which of those edges should form binding restrictions on the associated coefficients. If a restriction is binding, it requires the two coefficients to be jointed together at the optimum, e.g. the coefficient for FDP-Oblig and FDP-Facul are constrained to be equal. Thus, after estimating the model, one can examine the set of binding restrictions to see which groups were formed.

Formally, structured sparsity encodes these restrictions via penalized maximum likelihood. It begins by creating a vector of coefficients ($\boldsymbol{\beta}$ where β_i represents an indicator variable for unit i, e.g. FDP-Oblig) that represent the effect for each party-referendum type combination; if there are p units, then $\boldsymbol{\beta}$ has p levels.[9] This model usually consists of other, unpenalized, variables; these

[9]It is also possible to write this with respect to some baseline, e.g. the coefficient is the change versus some

can be added without any complication. The un-penalized model (i.e. the agnostic approach with indicator variables) can be estimated by maximum likelihood, with some likelihood function ℓ that depends on the observed data $\boldsymbol{X}$ and outcomes $\boldsymbol{y}$.

$$\hat{\boldsymbol{\beta}}_{MLE} = \arg\max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y}) \tag{3}$$

Structured sparsity is induced by adding some penalty on $\boldsymbol{\beta}$ to Equation 3 that encourages elements of $\boldsymbol{\beta}$ to be set equal to each other (e.g. 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; Hastie, Tibshirani, and Wainwright 2015). The LASSO is 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, 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: $|\beta_i - \beta_j|$.[10] 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 putting a penalty on the linear combination of some elements of $\boldsymbol{\beta}$: If $\boldsymbol{d}_k \in \mathbb{R}^{p\times 1}$, 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 $\boldsymbol{d}_k^T\boldsymbol{\beta} = \beta_i - \beta_j$. Returning to the simplified example in Figure 1, Table 1 outlines the linear restrictions that are needed for each of the types of structure shown above. Each row corresponds to one link between two nodes in Figure 1 where the coefficients in the corresponding $\boldsymbol{d}_k$ vector are shown in the left panel. The right panel indicates which structures include that restriction. For example, the first row of the table ($\boldsymbol{d}_1^T\boldsymbol{\beta}$) corresponds to the difference between $\beta_{\text{CVP}-\text{Init}}$ and

baseline referendum i. The following discussion proceeds with only minor changes.

[10] Note, however, that unless the $\boldsymbol{D}$ matrix defined in Equation 4 has full row rank, the problem cannot be cast as simply a regular sparsity problem (Duan et al. 2016).

$\beta_{\text{CVP}-\text{Oblig}}$ and would be included in the fully connected, lattice, and party connected structures but *not* the referendum structure as those two groups do not share a referendum type.

Table 1: Penalties Corresponding to Different Structures

Coefficient for Heterogeneous Effect (β)							Structure			
·	CVP Init	CVP Oblig	CVP Facul	SVP Init	SVP Oblig	SVP Facul	Full	Lattice	Ref.	Party
$\boldsymbol{d}_1$	1	-1	0	0	0	0	✓	✓		✓
$\boldsymbol{d}_2$	1	0	-1	0	0	0	✓	✓		✓
$\boldsymbol{d}_3$	1	0	0	-1	0	0	✓	✓	✓	
$\boldsymbol{d}_4$	1	0	0	0	-1	0	✓			
$\boldsymbol{d}_5$	1	0	0	0	0	-1	✓			
$\boldsymbol{d}_6$	0	1	-1	0	0	0	✓	✓		✓
$\boldsymbol{d}_7$	0	1	0	-1	0	0	✓			
$\boldsymbol{d}_8$	0	1	0	0	-1	0	✓	✓	✓	
$\boldsymbol{d}_9$	0	1	0	0	0	-1	✓			
$\boldsymbol{d}_{10}$	0	0	1	-1	0	0	✓			
$\boldsymbol{d}_{11}$	0	0	1	0	-1	0	✓			
$\boldsymbol{d}_{12}$	0	0	1	0	0	-1	✓	✓	✓	
$\boldsymbol{d}_{13}$	0	0	0	1	-1	0	✓	✓		✓
$\boldsymbol{d}_{14}$	0	0	0	1	0	-1	✓	✓		✓
$\boldsymbol{d}_{15}$	0	0	0	0	1	-1	✓	✓		✓

Note: This table shows the various linear combinations that are imposed for the structures (fully connected, lattice, referendum, and party connected) outlined in the main text. A ✓indicates that the corresponding linear combination ($\boldsymbol{d}_k$) is part of the penalty corresponding to that structure. For example, $\boldsymbol{d}_1$ states that a penalty term of $|\beta_{\text{CVP}-\text{Init}} - \beta_{\text{CVP}-\text{Oblig}}|$ is included in the fully connected, lattice and party structures.

From there, the penalty can be formed by adding together the absolute value of each of the relevant restrictions, i.e. $|\boldsymbol{d}_1^T\boldsymbol{\beta}| + |\boldsymbol{d}_2^T\boldsymbol{\beta}| + \cdots |\boldsymbol{d}_{15}^T\boldsymbol{\beta}|$ for the fully connected structure. More compactly, the entire penalty can be represented by a matrix $\boldsymbol{D} \in \mathbb{R}^{K\times p}$ and the ℓ_1 norm of this times the coefficient vector ($\boldsymbol{\beta}$). The strength of the penalty, i.e. the amount of encouragement for the restrictions to bind, is determined by λ. The choice of λ will be discussed shortly.

$$\hat{\boldsymbol{\beta}}_{\text{SSparse}} = \arg\max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y}) - \lambda||\boldsymbol{D}\boldsymbol{\beta}||_1 \tag{4}$$

Different choices of $\boldsymbol{D}$ can lead to different types of penalization. I focus today on grouping together elements of $\boldsymbol{\beta}$.[11] However, the theoretical results, inferential framework, and associated

[11]This is sometimes known as "trend filtering on a graph" (Wang et al. 2016), graph-guided fused LASSO (Chen et al. 2012), "graph fused LASSO" (Tansey et al. 2017), amongst others. For other uses of structured sparsity via the generalized LASSO, see Tibshirani and Taylor (2011) or Ali and Tibshirani (2019).

software outlined shortly apply to any choice of $\boldsymbol{D}$ and thus provides a general framework for future applications.[12]

4 Quantifying Uncertainty For Structured Sparsity

Given the above penalized maximum likelihood framework, it is possible to estimate heterogeneous effects by grouping together elements of $\boldsymbol{\beta}$ based on prior knowledge of permissible groupings and the observed data. A key question, however, is how to quantify the uncertainty in those estimated effects. This would allow us to understand both whether a given effect is statistically significant as well as the probability that two units are given a similar effect. Unfortunately, the non-differentiable penalty and the fact that it induces sparsity complicates traditional approaches. The former means that it is not possible to simply examine the Hessian of the penalized log-likelihood to estimate uncertainty; the latter means that procedures such as the bootstrap will likely fail to correctly quantify uncertainty (see, e.g., Leeb and Pötscher 2005; Kyung et al. 2010 for discussion in the case of standard sparse models).[13]

To address this, a variety of scholars have turned to Bayesian methods to analyze traditional sparse models to quantify uncertainty (e.g. Kyung et al. 2010). This paper extends structured sparsity for a similar purpose and thus allows for appropriate quantification of uncertainty. I do this by exponentiating the penalty in Equation 4 ($-\lambda||\boldsymbol{D}\boldsymbol{\beta}||_1$) and defining this as the kernel of the prior on $\boldsymbol{\beta}$. Unlike traditional choices of prior that focus on being weakly informative, this choice is explicitly designed to be informative as that is required to induce sparsity. While others have used specific choices of $\boldsymbol{D}$ (e.g. Kyung et al. 2010; Roualdes 2015; Betancourt, Rodríguez, and Boyd 2017; Tansey et al. 2017; Faulkner and Minin 2018), this paper is the first to explore the theoretical properties of the general case of structured sparsity. First, I explicitly analyze conditions when the structured sparse prior and resulting posterior is proper. This extends existing research that does not explicitly examine posterior propriety from an improper structured sparse prior. Second, I demonstrate how to perform posterior inference given this prior—both to sample the full posterior

[12]The full definition of structured sparsity, discussed in Appendix A, adds an additional penalty to Equation 4 based on penalizing a quadratic combination of coefficients. I do not explore this in detail here, but note that it can be easily added to the inferential framework by using results in Kyung et al. (2010).

[13]If the heterogeneity (e.g. $\boldsymbol{\beta}$) is merely nuisance, one could adapt techniques from elsewhere in machine learning (e.g. Tibshirani et al. 2004; Chernozhukov et al. 2018) for generating a measure of uncertainty on some other key quantity of interest.

and quantify uncertainty and to estimate the posterior mode using a novel EM algorithm. This extends existing work on sparse Bayesian inference as well as providing an alternative method to estimate the penalized maximum likelihood for linear and non-linear models. Finally, I discuss extensions including the choice of λ and using other penalties to induce structured sparsity besides the LASSO.

4.1 Theoretical Analysis of Bayesian Structured Sparsity

Existing work using Bayesian structured sparsity does not examine the conditions for posterior propriety. Rather, most papers using applications of Bayesian structured sparsity note that the prior is improper and adjust it in some way to ensure propriety (e.g. Betancourt, Rodríguez, and Boyd 2017; Faulkner and Minin 2018). This both may be un-necessary but also distorts (slightly) the theoretically motivated pattern of sparsity.[14] Broadly, the lack of results on posterior propriety is important to rectify as it means that unless researchers carefully modify their choice of structure sparsity, an improper posterior could result. As the Gibbs Sampler will still seemingly function properly in those cases (Hobert and Casella 1996), this poses a danger to reliable inference. I thus derive a series of simple and easy tests for posterior propriety in the following theorems. First, if the prior is proper, this addresses those concerns automatically. Theorem 1 states that only if $\boldsymbol{D}$ is full column rank ($\text{rank}(\boldsymbol{D}) = p$) does it effectively constrain all elements of $\boldsymbol{\beta}$ and thus makes a proper prior. It also shows how the normalizing constant, in the case of an improper prior, can be factored into one part that depends on λ and one part that does not. Thus, this makes Bayesian inference on λ easily tractable.

Theorem 1 *For any $\boldsymbol{D} \in \mathbb{R}^{K \times p}$, the prior on $\boldsymbol{\beta} \in \mathbb{R}^p$ corresponding to a posterior mode with structured sparsity can be expressed as follows, where $w_{\boldsymbol{D}}$ is a constant that depends only on $\boldsymbol{D}$ and not λ. The prior is proper, i.e. $w_{\boldsymbol{D}}$ is finite, if and only if* $\text{rank}(\boldsymbol{D}) = p$.

$$p(\boldsymbol{\beta}) = \lambda^{\text{rank}(\boldsymbol{D})} w_{\boldsymbol{D}} \exp\left(-\lambda ||\boldsymbol{D}\boldsymbol{\beta}||_1\right) \tag{5}$$

Note that for work on traditional sparsity (i.e. the Bayesian LASSO; Park and Casella 2008),

[14] It is telling that only rarely is the structured sparse prior proper by choice of $\boldsymbol{D}$ (e.g. Kyung et al. 2010's used of the fused LASSO – although note this shrinks not only to adjacent values but also to zero). Tansey et al. (2017) restricts inference to a proper subspace (e.g. p. 1053).

propriety of the prior follows automatically as $\boldsymbol{D} = \boldsymbol{I}$. Unfortunately, for most cases of structured sparsity, $\boldsymbol{D}$ is not usually full column rank. This occurs if there are any un-penalized variables, as is common, but also for many particular structures. Thus, it is important to know when the posterior is proper in all circumstances. Theorem 2 derives two simple and easily testable conditions for posterior propriety. Appendix B contains the proof, but the intuition is the following: Using an orthogonal rotation of the coefficients ($\boldsymbol{\beta}$) based on a singular value decomposition of $\boldsymbol{D}$, the prior factorizes into a proper structured sparse prior on a vector of length rank($\boldsymbol{D}$) and a flat prior on the other $p - \text{rank}(\boldsymbol{D})$ elements. With that in hand, existing results on the propriety of models with a (partially) flat prior (especially Michalak and Morris 2016) are employed.

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:* $p(\boldsymbol{\beta}) \propto \exp\left(-\lambda||\boldsymbol{D}\boldsymbol{\beta}||_1\right)$ *where* $\boldsymbol{D} \in \mathbb{R}^{K\times p}$

Consider the following two conditions, where $\mathcal{N}(\boldsymbol{A})$ *denotes the null space of a matrix* $\boldsymbol{A}$.

(a) Distinct Null Spaces of $\boldsymbol{D}$ *and* $\boldsymbol{X}$*: Either of the following, equivalent, conditions hold.*

$$\mathcal{N}(\boldsymbol{D}) \cap \mathcal{N}(\boldsymbol{X}) = \{\boldsymbol{0}\}$$
$$\text{rank}\left(\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{D} \end{bmatrix}\right) = p$$

(b) Unique MLE of Maximally Sparse Model: $\hat{\boldsymbol{\beta}}_{\mathcal{N}(\boldsymbol{D})}$*, defined below in two equivalent ways, exists and is unique.*

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

The following conditions characterize posterior propriety:

1. *(a) is necessary for posterior propriety.*

2. (b) is sufficient for posterior propriety.

The necessary condition (Condition a) states that the model must be either full rank in the design matrix ($\boldsymbol{X}$), have a proper prior (i.e. full column rank in $\boldsymbol{D}$), or that stacking them together creates an augmented design matrix that is full rank. The sufficient condition (Condition b) can be interpreted as follows: If one were to have a maximally sparse model (i.e. all restrictions in $\boldsymbol{D}$ were binding), if *that* model has a unique and finite MLE, then the posterior must be proper. In the case of the Fully Connected and Lattice structures, this is equivalent to checking whether a model with no heterogeneous effects (i.e. a single effect for each variable) is identified and has a finite MLE. Fortunately, this can be made even sharper for commonly used models; Corollary 1 leverages results on the connection between the existence of the MLE and posterior propriety in the (standard) multinomial case (e.g. Speckman, Lee, and Sun 2009) to draw out the following implications:[15]

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.*

Thus, for the models employed in this paper, there is a simple set of conditions to check beforehand to ensure that the posterior is proper. Researchers can thus safely use Bayesian structured sparsity while addressing concerns about an improper posterior.

4.2 Estimating Models with Bayesian Structured Sparsity

After establishing the key theoretical properties of structured sparsity, the next step is to show that this model is tractable. Theorem 3 does this by showing that, for a proper prior, it can be sampled easily using a Gibbs Sampler. It slightly extends results on Bayesian sparsity to the case of an arbitrary $\boldsymbol{D}$ (e.g. Andrews and Mallows 1974; Park and Casella 2008; Kyung et al. 2010); Tansey et al. (2017) provide a similar result.

Theorem 3 *For any $\boldsymbol{D} \in \mathbb{R}^{K \times p}$ that is full column rank, the corresponding structured sparse prior*

[15]For the linear model, Condition (a) implies Condition (b) and thus (a) alone is necessary and sufficient for propriety.

on $\boldsymbol{\beta} \in \mathbb{R}^p$ can be sampled by the following Gibbs Sampler.

$$1/\tau_k^2 | \boldsymbol{\beta}, \lambda \sim \text{InverseGaussian}\left(\frac{\lambda}{|\boldsymbol{d}_k^T \boldsymbol{\beta}|}, \quad \lambda^2\right) \quad \forall k \in \{1, \cdots, K\} \tag{8a}$$

$$\boldsymbol{\beta} | \{\tau_k^2\}_{k=1}^K \sim N\left(\mathbf{0}, \Sigma_\tau^{-1}\right); \quad \Sigma_\tau = \sum_{k=1}^K \frac{\boldsymbol{d}_k \boldsymbol{d}_k^T}{\tau_k^2} = \boldsymbol{D}^T \text{diag}\left(\{1/\tau_k^2\}_{k=1}^K\right) \boldsymbol{D} \tag{8b}$$

This simple representation clarifies a few points; first, in the simple Bayesian sparse case ($\boldsymbol{D} = \boldsymbol{I}$), the marginal prior on $\{\tau_k^2\}$ is the product of independent priors on each τ_k^2 and thus the joint distribution can be sampled exactly. Unfortunately, this property only holds when $rank(\boldsymbol{D}) = K$ (i.e. it has full row rank).[16] Second, as Appendix C derives in detail, this shows that a data augmentation scheme can be used to sample from the posterior for linear models as the conditional posterior on $\boldsymbol{\beta}$ given $\{\tau_k^2\}$ is normally distributed. Through various data augmentation schemes (e.g. Albert and Chib 1993; Polson, Scott, and Windle 2013), it is possible to create a simple Gibbs Sampler for binary (logistic and probit), multinomial logistic, and negative binomial regression. Thus, most common types of outcomes can be tractably analyzed using Bayesian structured sparsity.

While the key benefit of the Bayesian approach is to quantify uncertainty in the coefficients by sampling from the entire posterior distribution, its major downside is that it may be slow for sufficiently large datasets as well as the fact that there is not exact sparsity in the sampled posterior. As noted before, the posterior mode (corresponding to the traditional penalized MLE) is of important substantive interest as it will be (structured) sparse and contain exact groups of observations. Thus, it is important to have a method for estimating that posterior mode for an arbitrary $\boldsymbol{D}$ and for non-linear models. Unfortunately, directly maximizing that objective (Equation 4) is difficult because the penalty is not differentiable, and thus standard optimization methods will fail (Tseng 2001). Many different methods have been proposed to solve this (e.g. Zhu 2017; Gaines, Kim, and Zhou 2018; Chen et al. 2012; Oelker and Tutz 2017; Tansey et al. 2017). Unfortunately, these methods are more tailored to specific questions versus a general, possibly non-linear, scenario; for example, many methods exist based on Alternating Direction Method of Multipliers.[17] While

[16] This interestingly mirrors a key result on the non-Bayesian generalized LASSO, e.g. Duan et al. 2016: If $rank(\boldsymbol{D}) = K$, the generalized LASSO can be transformed exactly into a standard LASSO problem. In all other cases, it becomes a LASSO problem with equality constraints (e.g. Gaines, Kim, and Zhou 2018).

[17] Zhu (2017) and Gaines, Kim, and Zhou (2018) for any $\boldsymbol{D}$; Tansey et al. (2017) for the class of $\boldsymbol{D}$ employed in

having simple closed form solutions for linear models, they require additional modification to efficiently solve non-linear models. Proximal methods (e.g. Chen et al. 2012) are more promising in being able to target non-linear likelihoods directly, but existing work is tailored to specific choices of $\boldsymbol{D}$. Broadly, it does not appear that existing methods can easily and exactly handle both (i) non-linear models and (ii) an arbitrary choice of $\boldsymbol{D}$.

A by-product of the Bayesian approach is that it provides a simple way to address these concerns for any $\boldsymbol{D}$ with a linear or multinomial likelihood. Appendix C derives a novel EM algorithm to locate the posterior mode (i.e. the traditional structured sparse estimate) using only iterative least squares for linear and multinomial likelihoods. The algorithm has the usual advantages of an EM procedure (e.g. guaranteed to increase the likelihood at each step; lacks internal tuning parameters; see McLachlan and Krishnan 2008 for an overview).[18] While using an EM approach is sometimes used for models with traditional sparsity (e.g. Figueiredo 2003; Polson and Scott 2011b; Ratkovic and Tingley 2017), it is typically slower and thus the methods outlined above are preferred. In the case of structure sparsity, however, Appendix G shows that the EM algorithm the associated software is often noticeably faster—up to an order of magnitude in some cases—than alternatives (e.g. Zhu 2017) as the size of the data grows. As this algorithm exactly targets the posterior mode—without approximation—and works for a general penalty, it is useful for estimating structured sparsity with non-linear models on huge datasets beyond where Bayesian methods are feasible. Other applications include accelerating Bayesian methods by providing good starting values, allowing exact analysis of the posterior mode to interpret groups, and for tasks involving predictive inference where uncertainty is not required.

4.3 Extensions of Bayesian Structured Sparsity

The above discussion has assumed two things that can be easily relaxed. First, it has assumed that λ is fixed by the researcher. Second, it assumed that a LASSO-type penalty (i.e. an ℓ_1 penalty on $\boldsymbol{D}\boldsymbol{\beta}$) was the best way to induce structured sparsity. This section brief outlines how each can be relaxed.

this paper.

[18] Note that the EM algorithm is only guaranteed to find a local mode. However, in many cases, the posterior mode given a structured sparse prior is unique and thus the EM algorithm will find the unique global mode, given starting values that are not at fixed points. See Ali and Tibshirani (2019) for detailed discussion of when uniqueness holds.

The choice of λ is important as it governs the amount of sparsity in the model; as it goes to zero, the effect of the prior goes away—and the model is not sparse (e.g. the agnostic approach of interacted indicator variables). As it goes to infinity, the model becomes one that is fully sparse (e.g. all restrictions are binding; $\boldsymbol{D\beta} = \mathbf{0}$). There is a tradeoff when selecting this variable; less sparse models are more complex and fit the observed data better, but incur higher variance and may "overfit" to the particular data observed. On the other hand, sparser models fit less well and thus may incur higher bias. This bias-variance tradeoff is central in machine learning and is discussed extensively elsewhere, e.g. Hastie, Tibshirani, and Friedman (2009, ch. 7). There is a vast literature dedicated to how to select the right point on that trade-off in a general sense and for specific models. Broadly speaking, researchers are able to use their preferred method when applying structured sparsity, but I briefly outline some general points that are relevant to approaches based on methods other than standard cross-vadliation. Appendix D provides a more in-depth discussion.

First, a common way to select models involves using an information criterion such as the AIC or BIC. That requires evaluating the log-likelihood as well as a measure of complexity of the model. Tibshirani and Taylor (2012) provide an unbiased measure of the degrees of freedom for the generalized LASSO in the linear model that can be used to calculate these information criteria. Second, an advantage of the Bayesian approach is that it facilitates non-nested model comparison using criterion such as the WAIC (Gelman, Hwang, and Vehtari 2014; Vehtari, Gelman, and Gabry 2017) that are designed to approximate cross-validation but do not require sample splitting. Finally, one can use a standard Bayesian approach and set a prior on λ and thus allow it to be also estimated from the model. This somewhat begs the question insofar as calibrating the prior on λ is a similarly difficult task; a weakly informative prior may be inappropriate as the point of structured sparsity is to ensure that the prior binds. Thus, it is typically good practice to allow the prior mean to grow with the size of the data (see Ratkovic and Tingley 2017 for a similar idea in a traditional sparse case). Exploring whether a good plug-in value for λ can be established for structured sparsity is an interesting area for future research.

In terms of using a penalty other than the LASSO, a large literature has developed to replace the LASSO-type penalty (i.e. the ℓ_1 norm) with other ways to create sparse models. In the Bayesian literature, a popular form of sparsity is known as global-local models (Polson and Scott 2011a). These models can be cast as scale mixture of normal distributions, i.e. a mixture on the

variance component of the normal distribution. This includes the LASSO as a particular case, but also many other common priors such as the adaptive LASSO, horseshoe, double Pareto, and LASSOplus (respectively, Zou 2006; Carvalho, Polson, and Scott 2010; Armagan, Dunson, and Lee 2013; Ratkovic and Tingley 2017). These priors are designed to address issues in the LASSO, i.e. finite sample bias in the non-zero coefficients, and often have an "oracle" property where the posterior mode will converge to the estimates of a model fit only on the true (unknown) non-zero groups (Fan and Li 2001; Zou 2006, but see Leeb and Pötscher 2005 for a cautionary note). For structured sparsity, a similar oracle result based on the adaptive LASSO has been shown for a fully connected structure in a variety of settings Bondell and Reich (2009), Gertheiss and Tutz (2010), and Chen (2015).[19]

Implementing these other global-local priors can be done easily using structured sparsity; Theorem 4 states the result—Theorems 1 and 2 govern posterior propriety and the full conditionals for the augmentation variables and the coefficients are in the same family as in the standard sparse case.Note that if the distribution on τ_k^2 was $\text{Exp}(\lambda^2/2)$, the original result in Section 4 is recovered. One complication arises concerning inference on the hyper-parameters ($\boldsymbol{\lambda}$). Given that the marginal distribution on $\{\tau_k^2\}_{k=1}^K$ is intractable, work is needed on a case-by-case basis to factorize the normalizing constant (i.e. pulling out the analogous term to $\lambda^{\text{rank}(\boldsymbol{D})}$ in Theorem 1). If that is not possible, using methods such as doubly-intractable sampling may prove necessary (e.g. Park and Haran 2018).

Theorem 4 *Assume that the prior on δ is a global-local prior (Polson and Scott 2011a) whose marginal density $p_{g,\boldsymbol{\lambda}}(\delta)$ can be expressed as follows, where $\boldsymbol{\lambda}$ is a fixed vector of hyper-parameters and g is some proper probability distribution whose support is on the non-negative reals:*

$$p_{g,\boldsymbol{\lambda}}(\delta) = \int_0^\infty (2\pi\tau^2)^{-1/2} \exp\left(-\frac{\delta^2}{2\tau^2}\right) g(\tau^2;\boldsymbol{\lambda}) d\tau^2$$

Define a global-local structured sparse prior as having the following (possibly improper) joint density:

[19] A similar proof can be derived for a general $\boldsymbol{D}$.

$$p\left(\boldsymbol{\beta}, \{\tau_k^2\}_{k=1}^K | \boldsymbol{\lambda}\right) \propto \exp\left(-\sum_{k=1}^{K} \frac{\boldsymbol{\beta}^T \boldsymbol{d}_k \boldsymbol{d}_k^T \boldsymbol{\beta}}{2\tau_k^2}\right) \prod_{k=1}^{K} \frac{g(\tau_k^2; \boldsymbol{\lambda})}{\sqrt{\tau_k^2}}$$

The following properties characterize the marginal prior on $\boldsymbol{\beta}$ implied by this joint prior:

1. *The prior on $\boldsymbol{\beta}$ is proper if and only if* $\text{rank}(\boldsymbol{D}) = p$.

2. *Theorem 2 characterizes posterior propriety for any permissible g.*

3. *If proper, the prior can be sampled where $\{\tau_k^2\}_{k=1}^K | \boldsymbol{\beta}$ is in the same family as in the standard sparse case (i.e.$\boldsymbol{D} = \boldsymbol{I}$) and $\boldsymbol{\beta} | \{\tau_k^2\}_{k=1}^K$ is normally distributed.*

5 Simulations on Estimating Heterogeneous Effects

I begin by examining how existing methods for estimating heterogeneous effect fare when the underlying heterogeneity is based on *groups* of distinct values. I employed the simple simulation environment below; it differs from existing approaches insofar as there are G units for which effects are to be estimated but the true heterogeneity is quite simple insofar as most units have an effect of either '1' or '-1'. 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. The conjecture underlying the paper is that existing methods based on shrinking towards a global effect will do less well in this scenario than structured sparsity which is explicitly designed to estimate groups of distinct effects.

- 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'.

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

- Outcome:
  - The generative model is as follows:

$$y_i = x_i + \tau_{g[i]} d_i + \epsilon_i; \quad x_i \sim N(0,1); \quad \epsilon_i \sim N(0,1)$$

  - Half of the units in each group $r/2$ are in treatment ($d_i = 1$) and half are in control ($d_i = 0$) at random.
- 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. The full formula and hyper-parameter choices are outlined in Appendix E.

Following existing simulations for heterogeneous effects, I use these models to estimate the effect of d_i (from zero to one) for each unit g, marginalizing over x_i and compare that estimated effect against the true value.[20] This fits into existing traditions that seek to estimate heterogeneous effects by correctly estimating the conditional mean of the outcome (e.g. $E[y_i|d_i, g[i]]$) and thereby back-out the implied treatment effects.[21]

Appendix E compares a wide array of methods, with eight key approaches reported here.[22] Table 2 shows the RMSE (root mean squared error) of the estimated sub-group effects versus their true value, averaged across one-hundred simulations. It also varies the number of observations per group ($r \in \{10, 20, 50, 100\}$) to see how the models performed with more data. Finally, to see the effect of the group structure, I ran simulations where almost all groups had a non-zero effect ($\mathcal{S} = \lfloor G/2 \rfloor$, i.e. $G/2$ if G is even) and other, more traditional, simulations where roughly half of

[20] The true effect is calculated for the linear case by the relevant τ_g. For some methods, the effect depends on the observed x_i. I address this by calculating the marginal conditional average treatment effect (MCATE; Grimmer, Messing, and Westwood 2017) for all methods. I do this using Monte Carlo integration : I draw 1,000 observations from a standard normal distribution: $\{\tilde{x}_i\}_{i=1}^{1000}$. For each group g, I calculate the individual estimated effect for each observation i: $E[y_i|d_i = 1, g[i] = g] - E[y_i|d_i = 0, g[i] = g]$. Averaging those together gets the estimate of the MCATE for each method and group.

[21] These simulations can thus be seen as the ability to recover a high-dimensional conditional expectation function correctly.

[22] This follows the methods presented in Grimmer, Messing, and Westwood (2017), with the exception of KRLS (Hainmueller and Hazlett 2014) that proved too memory intensive to fit for the large datasets.

the groups had a *zero* effect ($\mathcal{S} = \lfloor G/4 \rfloor$).

Table 2: Estimating Heterogeneous Effects for $G = 25$ Units

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

r	SSp (AIC)	FE	RE	BayesGLM	LASSO	ENet 1	FindIt	BART
10	0.377	0.450	0.429	0.448	0.575	0.568	0.447	0.971
	(0.008)	(0.006)	(0.007)	(0.006)	(0.008)	(0.007)	(0.007)	(0.001)
20	0.239	0.312	0.307	0.311	0.400	0.410	0.307	0.918
	(0.005)	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.002)
50	0.143	0.202	0.208	0.202	0.274	0.285	0.200	0.548
	(0.004)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.006)
100	0.095	0.140	0.142	0.140	0.184	0.192	0.138	0.288
	(0.003)	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.004)

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

r	SSp (AIC)	FE	RE	BayesGLM	LASSO	ENet 1	FindIt	BART
10	0.412	0.448	0.397	0.445	0.498	0.486	0.444	0.690
	(0.007)	(0.007)	(0.006)	(0.007)	(0.007)	(0.006)	(0.007)	(0.002)
20	0.294	0.314	0.301	0.313	0.380	0.381	0.325	0.653
	(0.005)	(0.005)	(0.004)	(0.005)	(0.005)	(0.005)	(0.004)	(0.002)
50	0.167	0.200	0.201	0.200	0.289	0.293	0.204	0.433
	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.005)
100	0.111	0.143	0.146	0.143	0.223	0.227	0.143	0.238
	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)

Note: The table reports the RMSE for the estimated heterogeneous effects versus the true values in $\boldsymbol{\tau}$ averaged across 100 simulations. The standard error of the RMSE is shown in parentheses. r refers to the number of observations per group; half are treated, half are control - allocated at random. The top panel shows a simulation where most $\boldsymbol{\tau}$ are not zero (i.e. $\lfloor G/2 \rfloor = 12$ are '1', 12 are '-1' and one is zero). The later shows a case where the majority are zero (i.e. 13 are zero; 6 are '1' and 6 are '-1').
The methods used are outlined in detail in Appendix E but are, in order, structured sparsity with λ chosen via the AIC (SSp [AIC]); an interactive model (fixed effects - FE); a hierarchical model with random slopes (RE); a regression with a weakly informative prior from Gelman et al. 2008 (BayesGLM); LASSO with λ chosen via ten-fold cross-validation; Elastic Net with $\alpha = 0.5$ (ENet1); FindIT (Imai and Ratkovic 2013) using the default settings; Bayesian Additive Regression Trees fit using `BART` (Sparapani, Spanbauer, and McCulloch).

It shows 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, i.e. varying the number of observations per group (r), structured sparsity performs the best. It has a smaller RMSE than all other methods even when the number of observations is large; further, its advantage remains even when the dataset is of a small size (e.g. $r = 10$). This confirms that it can effectively exploit group-structured heterogeneity when that well-describes the data generating process.

It is also worth noting how other methods perform; a key point to note is that methods based regression trees do rather poorly in these simulations (e.g. BART, same for random forests—see Appendix E). This makes sense given the nature of tree-based methods; they prevent over-fitting by limiting the depth of each individual tree. However, as the only way to pull some group g from the global effect is to separate it via a split on a single indicator variable, deep trees are required to accurately capture all of the sub-group effects. Since such deep trees are discouraged, these methods will systematically underestimate the sub-group effects.[23]Similarly, methods based on traditional sparsity (LASSO and the elastic net [ENet1]) also do rather poorly. Given they are based on assuming that most of the treatment-group interactions are zero, it will struggle to recover the correct structure. FindIt (Imai and Ratkovic 2013) does comparably well—usually in about second place—but interestingly does about the same as a more simple hierarchical random effects model.

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'. In this case, structured sparsity does worse than before. While it still is the best-performing method as long as the ratio (r) is moderately high, its RMSE is higher than before and more narrowly out-performs other methods. While tree-based methods still do rather poorly, the simulations show (as expected) that LASSO-based methods (LASSO and elastic net [ENet 1]) do noticeably better in this second environment given that their fundamental assumption (standard sparsity) remains satisfied.

Appendix E provides further exploration of these effects by varying two aspects of the simulation. First, it varies the number of units G widely ($G \in \{5, 10, 25, 50, 100\}$). With the exception of $G = 5$ where structured sparsity effectively mirrors an interactive (fixed effects) based approach, the same pattern remains: Structured sparsity out-performs all methods for a mostly grouped data generating

[23]There is a careful point to note about how trees treat categorical variables. In the most straightforward sense, the identifiers for each group are included as a series of binary variables that are treated like any other input to a tree. However, some packages have a different way of treating categorical variables (e.g. `randomFoest`; Liaw and Wiener 2002): Rather than including them as indicator variables, they actually order the categories at each split based on the observed outcome and use that *ordered* variable to partition groups (see, e.g., Hastie, Tibshirani, and Friedman 2009, p. 310). It is unclear whether this is appropriate for an inferential (i.e. non-predictive) task as it, in some sense, uses the outcome to create variables to help predict the outcome! I thus report results from a "fair" random forest that includes these variables as a series of indicators; the "unfair" random forest is shown in Appendix E. It does noticeably better—sometimes being the best performing method, especially as the number of groups G grows large (e.g. $G = 100$).

process ($\mathcal{S} = \lfloor G/2 \rfloor$) and usually is best-performing methods when the data generating process is mostly sparse ($\mathcal{S} = \lfloor G/4 \rfloor$) and with a moderate number of observations per group ($r > 10$).

Second, I examined a model with a binary outcome and a logistic data generating process. In this case, the broad pattern remains; in the grouped data generating process, structured sparsity performs well compared to other methods, especially as r increases. For modest r (e.g. $r = 20$), it is usually out-performed by a hierarchical random effects model (RE) while still beating most other methods. When the data generating process is more sparse ($\mathcal{S} = \lfloor G/4 \rfloor$), it performs in the middle of the set of methods examined. Exploring why the performance degrades in a non-linear setting is worthy of further exploration; I conjecture this is because model selection using the AIC or BIC no longer has an unbiased estimator of the degrees of freedom as the estimator in Tibshirani and Taylor (2012) is only unbiased for the linear model.Extending existing work to correct this measure for the standard sparse case (e.g. Ninomiya and Kawano 2016) to the generalized LASSO with non-linear outcomes is an important area for future research.

6 Uncovering Heterogeneity Amongst Swiss MPs

Returning to the empirical example discussed at the start of the paper, I use structured sparsity to uncover heterogeneous effects of interest group affiliation in Switzerland (Giger and Klüver 2016). Before turning to structured sparsity, it is necessary to re-specify their original results to address a potential theoretical issue.[24] A key theoretical frame for understanding an MP's decision-making process is known as "competing principals" (Carey 2007): Many actors make demands on MPs. The two most important principals are their constituent preferences and their parliamentary party.[25] Giger and Klüver (2016) add important nuance to this story by noting that interest group affiliations can be another source of influence on MPs and cause systematic defections against their constituents. They measure susceptibility to interest group pressure by the number of formal connections to economic or cause interest groups as measured by a legally required register of interests discussed in Section 2. Their theoretical expectations are that MPs with links to groups

[24]There is also a subtle point about temporal sequencing, noted by Barceló (2017), that the constituent preferences are revealed after the MPs vote. For the purposes of this paper, I focus on the original analysis by Giger and Klüver (2016).

[25]Even though party discipline in Switzerland is thought of as "moderate" (Kriesi and Trechsel 2008), following the leadership should still have some bite.

representing economic ("sectoral") interests will be pulled to defect from their local constituents while MPs with more ties to cause groups (e.g. NGOs) will be pulled more to agree with their local constituents. Unfortunately, they operationalize their explanatory and dependent variables in a way that renders their analysis difficult to interpret. Table 3 illustrates this by outlining the four logically possible outcomes, assuming without loss of generality that the local constituents vote "yes" on a particular issue.

Table 3: Examining Competing Principals

Scenario	Choice			Outcome	Implications		
	Local	Party	MP		Defect	Party Congruence	Cross Pressured
(1)	Yes	Yes	Yes	Follow Consensus		✓	
(2)	Yes	Yes	No	Reject Consensus	✓		
(3)	Yes	No	Yes	Local Over Party			✓
(4)	Yes	No	No	Party Over Local	✓	✓	✓

Note: This table shows the logically possible scenarios, conditional on the local canton choosing "yes". "Outcome" provides a concise label for how the MP's action should be interpreted. "Implications" outline relevant variables implied by this scenario.

From the competing principals framework, there are two distinct scenarios. First, MPs may be "cross-pressured" when their constituents and party disagree on a particular issue. In that case, the MP must decide to support one or the other—they cannot satisfy both sides. By Giger and Klüver (2016)'s logic, one would expect the same hypotheses to hold—economic groups pull MPs towards the party and cause groups pull towards the locality. A fundamentally different scenario occurs, however, when the MPs are not cross-pressured. When their party and canton agree, the MP has a simple and natural choice—follow both principals. Again, however, interest group pressure may cause defections from this consensus.

The core point is that a defection under those two different scenarios implies something very different; the first is a choice of party against the constituents, the second is a rejection of *both*. Unfortunately, existing analyses using this data (implicitly) conflates these two scenarios by mixing them together by only examining "defect from local constituents" without recognizing the two distinct meanings of this defection (Portmann, Stadelmann, and Eichenberger 2012; Stadelmann, Portmann, and Eichenberger 2013; Giger and Klüver 2016; Barceló 2017). This problem is exacerbated by the inclusion of an inappropriate control variable: "congruence with official party

position" (Giger and Klüver 2016; Barceló 2017). This variable measures whether the MP voted in accordance with their party's official position. Note, however, it actually depends on *both* the MP's decision and the *party*'s decision. Thus, the outcome one seeks to explain (a transformation of the MP's voting decision) appears, implicitly, as both an independent and dependent variable. This leads to a conceptual problem when thinking of the effect of other covariates.[26]

Fortunately, this problem can be easily rectified: Split the data into those votes by whether MPs are cross-pressured, and run the original specification from Giger and Klüver (2016), excluding the variable for party congruence. Table 4 does this and provides sharply different results. In the case of an MP being cross-pressured, their analysis is partially confirmed. MPs with more economic ties are more likely to side with their party over their constituent preferences when they disagree. However, there is no significant effect of cause interest groups, in either direction. In the case of an MP being *not* cross-pressured, the story is rather different. In this case, a defection involves defying both their party and their constituency; we see here, clearly, that *either* type of interest group attachment lowers the probability of defection. Thus, more connected MPs are more likely to support the consensus position (and agree with their locality) versus rebelling against both principals.

Table 4: Re-Analyzing MP Defection

	Replicating Giger and Klüver (2016)	Defect from Local	Defect from Consensus
Number of Economic Groups	0.013*	0.123*	−0.032*
	(0.005)	(0.016)	(0.012)
Number of Cause Groups	−0.035*	−0.009	−0.061*
	(0.008)	(0.019)	(0.020)
Observations	20,260	7,321	12,939

Note: Coefficients and standard errors reported. * indicates a p-value below 0.05. Models estimated using `glmer`. The first model exactly replicates the results from Giger and Klüver (2016). The second model ("Defect from Local") uses only the observations where MPs are cross-pressured (i.e. canton and party disagree). The third model ("Defect from Consensus") uses only the observations where there is a consensus position (e.g. the canton and party agree).

[26]Consider the case where the MP is assumed to be in congruence with their party; one can therefore interpret the coefficient on, say, interest group as the effect of that variable holding constant an MP's congruence with their party. A "defection" corresponds to "Party Over Local" (4, in Table 3). Non-defection does not correspond to the obvious "Local Over Party" but rather the choice of "Follow Consensus". Crucially, in comparing those two outcomes, *the party position must change*. There is no logically possible scenario where we can (a) hold constant the local preference, (b) hold constant an MP's congruence with their party, but (c) switch the outcome from "defection" to "non-defection".

6.1 Applying Structured Sparsity

I show how structured sparsity can identify important sub-group heterogeneity, i.e. whether a key explanatory variable has a different effect for sub-types of observations. All of the preceding analysis focused on the aggregate effect of interest group attachment. However, it is important to be able to understand which groups are driving an average effect while maintaining interpretability and understanding by relying on a grouped approach. I briefly outline two dimensions that I explore for heterogeneous effects, although there are clearly other aspects that could be explored.[27]

I first examine the coding of referenda into three distinct types as noted in the earlier exposition of structured sparsity. To provide more context, Giger and Klüver (2016) note the referenda fall into three distinct types that have very different temporal sequencing and origins. There are "obligatory" referenda; these are proposals to amend the Swiss constitution and must be ratified by a "double majority" (i.e. majority of the popular vote and a cantonal majority).[28] For these proposals, the MPs move first and pass the proposal than is then sent to the public for ratification. In the period studied here, 18 of 21 proposals are enacted (Stadelmann, Portmann, and Eichenberger 2013). Next are "facultative" referenda; these are citizen initiated referenda to *challenge* enacted legislation. In this case, public signatures or eight cantons can require a referendum to repeal existing legislation (Kriesi and Trechsel 2008). In this data, 30 of 37 votes upheld the enacted law (Stadelmann, Portmann, and Eichenberger 2013).[29] Finally, there are "initiatives". These have a rather different origin; they are *proposals* to modify the constitution that are sent to the legislature after a petition. The legislature then votes on the proposal to send a sense of its preference, but its vote is not binding as the proposal is voted on by the public.[30] They are much less successful with only 6 out of 60 in the data being enacted (Stadelmann, Portmann, and Eichenberger 2013).[31] Kriesi and Trechsel (2008) suggests that the rise of the right-populist Swiss People's Party (SVP)

[27] For example, Barceló (2017) explores the interaction between interest group attachment and the margin of victory of the referendum, although using the flawed coding scheme as outlined above. Carey and Hix (2012) re-analyses Portmann, Stadelmann, and Eichenberger (2012) to look for non-linear effects of district magnitude. Many other extensions are possible.

[28] Kriesi and Trechsel (2008, p. 39) define a getting a majority of votes cast in the following way: "every canton has one vote (which corresponds to the popular majority in each canton), with the six former half-cantons having only half a vote".

[29] This is a fairly high rate of acceptance compared to earlier periods, see Kriesi and Trechsel (2008, p. 58).

[30] This procedure has some additional complexities based on the type of proposal by the citizens and whether Parliament seeks to draft a counter-proposal; see Kriesi and Trechsel (2008, ch. 4) for a detailed discussion.

[31] This rate is higher than much of the 20th century, as only a single initiative was accepted between 1928 and 1982 (Kriesi and Trechsel 2008, p. 59).

was accompanied by a greater adoption of the initiatives by right-wing causes.

However, this may not be the only dimension of effect heterogeneity. It is also likely the case that there are different effects by the *party* of the MP. One could imagine differences arising from ideological backgrounds (e.g. how are links to interest groups viewed) or from the institutionalization of the parties (e.g. how connected are parties to the policy-making process). For example, there are four major parties that are continuously represented in the executive Federal Council.[32] A plausible conjecture is that interest group attachment may operate differently for MPs inside these major parties than for the numerous smaller parties in the legislature.

As outlined in Section 2, I will explore for heterogeneous effects by the interaction of party and referendum type. As there is a lack of clear prior theoretical expectations, structured sparsity provides a natural way to uncover if any sub-groups have systematically different effects. Given the large number of categories (sixty for each type of interest group), an agnostic approach is likely to run into serious difficulties. I thus use structured sparsity to stabilize the estimates of the sub-group effects. Using this method requires selecting a structure ($\boldsymbol{D}$) and a strength of sparsity (λ). I do this by using a combination of Bayesian cross-validation and theory; Appendix F provides the full details, summarized here: I fit a model across a grid of λ for the fully connected, lattice, and referendum connected structures outlined in Section 2 where the penalties are weighted by an adaptive LASSO-type penalty (Zou 2006; Gertheiss and Tutz 2010).[33] Examining the fit of the models, both the fully connected and lattice structures out-perform the referendum structure. As the fit of those two are very similar, I adopt a more parsimonious view and use the lattice structure for the main results presented here as it imposes some prior coherence on the estimated groups.

I next compare structured sparsity against three other models that are commonly used in an observational regression context—with random effects—for estimating sub-group heterogeneity. First, I replicate the original specification and estimate a model with only a single aggregate effect. Second, I estimate the fully interacted model with a diffuse normal prior on each interaction term to stabilize the model and ensure a proper posterior (i.e. a ridge penalty). Finally, I examine against a natural alternative approach: a normal hierarchical model (Gelman and Hill 2006). In addition to

[32] I use "major parties" to denote the parties included in the "magic formula" for permanent representation in the executive branch (Federal Council; see Kriesi and Trechsel (2008) for an overview): FDP (Free Democratic Party), SDP (Social Democratic Party), SVP (Swiss People's Party), and Christian Democratic People's Party (CVP).

[33] The party connected structure does not provide enough stabilization to result in a proper posterior by application of Corollary 1.

Table 5: Comparing Models for Heterogeneity

Cross-Pressured Votes				
	Model			
Statistic	No Heterogeneity	Interactive	Random Effects	Structured Sparsity
Log-Likelihood	-2803.19	-2657.1	-2684.2	-2673.23
Complexity	35.58	112.28	71.39	56.35
WAIC	5677.55	5538.75	5511.18	5459.17
Consensus Votes				
	Model			
Statistic	No Heterogeneity	Interactive	Random Effects	Structured Sparsity
Log-Likelihood	-3256.6	-3173.46	-3205.92	-3196.78
Complexity	27.58	95.96	50.28	42.62
WAIC	6568.36	6538.83	6512.4	6478.8

Note: These tables compare a Bayesian measure of model that looks at the in-sample fit ("Log-Likelihood") and adjusts by complexity of the model ("Complexity") to get a WAIC. Smaller numbers represent better model fit. The models are described in the main text.

a global effect, I add a random slope for the effect of each interest group by party-referendum type combination. I compare these (non-nested) models using the WAIC [Widely Applicable Information Criterion].[34] The WAIC can be interpreted analogously to a standard information criterion (e.g. AIC or BIC): It examines the fit of the model (in terms of the average log-likelihood) but penalizes more complex models. The penalty can be very roughly thought of as the "effective" number of parameters in the model. A smaller WAIC indicates a better performing model. Table 5 compares the selected model against the other specifications for addressing heterogeneity. Consider the comparison of the model with no heterogeneity to the agnostic interactive model. While the interactive model fits the observed data better (higher Log-Likelihood), it does so at a large price to complexity—nearly tripling the complexity penalty.

[34]Gelman, Hwang, and Vehtari (2014) or Vehtari, Gelman, and Gabry (2017) provide accessible expositions of the WAIC. Formally, noting that N is the number of observations, S is the number of posterior draws and $\boldsymbol{\theta}^{(s)}$ as the vector of parameters associated with draw s. $Var(\{a^s\}_{s=1}^S)$ indicates the sample variance of a set of observations $\boldsymbol{a}$ indexed by s.

$$\hat{lpd} = \sum_{i=1}^{N} \ln\left(\frac{1}{S}\sum_{s=1}^{S} p(y_i|\boldsymbol{\theta}^{(s)})\right)$$
$$\hat{p}_{WAIC} = \sum_{i=1}^{N} Var\left(\{\ln p(y_i|\boldsymbol{\theta}^{(s)})\}\right)$$
$$\hat{WAIC} = -2\hat{lpd} + 2\hat{p}_{WAIC}$$

Table 5 uses "Log-Likelihood" to refer to $\hat{lpd}$ and "Complexity" to refer to $\hat{p}_{WAIC}$.

Given the dramatic increase in model complexity using the interactive model, it is worth examining whether some regularization can estimate heterogeneous effects at a more reasonable cost of complexity. As Table 5 shows, they can: Both the random effects model and structured sparsity fit the data better than the original model, while imposing a more modest penalty in complexity and thus have better performance (smaller WAIC). When comparing these two models against each other, we see that structured sparsity outperforms the random effects models in both cases, although the difference is larger in the case of cross-pressured votes. It is interesting also to note that structured sparsity both fits the data better (higher log-likelihood) at a lower complexity than random effects in both cases. A heuristic test of whether this is statistically significant (Gelman, Hwang, and Vehtari 2014) suggests that it is (t-stat of 4.71 and 4.80 for cross-pressured and consensus votes, respectively).[35]

6.2 Examining Groups Using Bayesian Structured Sparsity

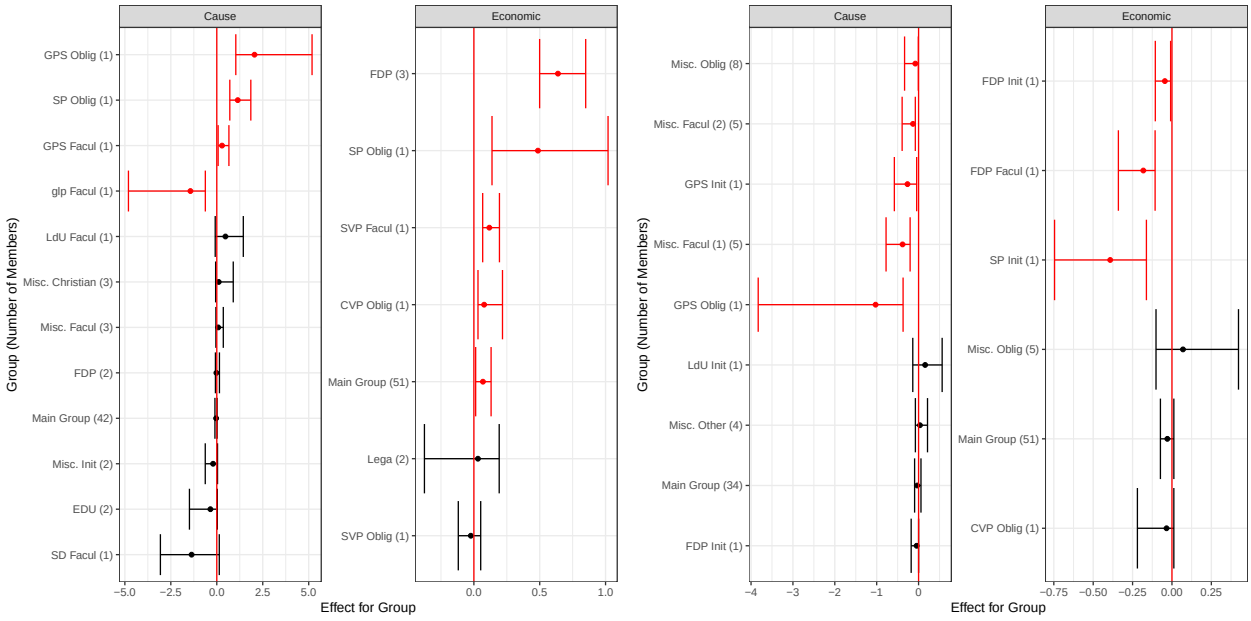
A key advantage of using Bayesian structured sparsity is that it facilitates interpretation of the estimated heterogeneous effects by looking at the groups that formed. Using the groups found at the posterior mode, I labelled each of them by their primary characteristic (see Appendix F for full listing). For all models (consensus, cross-pressured and economic, and cause), there was typically one large group that consisted of most of the sub-group effects. Interestingly, however, a number of small and distinct sub-groups emerged. Figure 2 shows the estimated effect for each sub-group. The point indicates the posterior mode estimate that all members of the group are given; to approximate uncertainty, I reported the average of the coefficients for units in that group across the posterior samples.

The figure shows a number of interesting points; first, the FDP stands out as clearly distinct from all other parties. Looking at economic interest groups, its three sub-group effects (for all referendum types) are put into a single group with a much larger effect than other parties for cross-pressured votes and two of the three sub-groups are distinct and significantly negative for consensus votes. Similarly, for cause groups on cross-pressured votes, two of the three effects are

[35] In both cases, structured sparsity is favored over the model with no heterogeneity (t-stat of 7.64 and 4.73, for cross-pressured and consensus votes respectively). The interactive model is statistically significantly favored over one with no heterogeneity in the cross-pressured case (t-stat of 3.37) but not in the case of consensus votes (t-stat of 0.96).

Figure 2: Estimated Groups Using Structured Sparsity

(a) Cross-Pressured Votes (b) Consensus Votes



Note: The groups located at the posterior mode for the optimal λ (see main text for a discussion) are shown here. Each panel shows the groups created for cause and economic interest groups, respectively. The model fit on cross-pressured votes is shown in the left figure (2a); the model on consensus votes is on the right. The dot indicates the value that all members in the group are assigned at the posterior mode. Uncertainty for each group is calculated by taking the average of the coefficients corresponding to the members over the entire sampled posterior. 95% credible intervals on that quantity are shown; intervals in red (and with a triangle marker) do not contain zero and are placed at the top of the figure.

located outside the main group—albeit with a much smaller effect. As the FDP is an economically liberal (i.e. pro-market) party, this could be explained by its closer attachments to economic groups and thus MPs who are more closely affiliated are more likely to defy their constituents in favor of the party line when they are cross-pressured. When there is a consensus vote (i.e. the party and constituents agree), those MPs with closer ties to the interest groups are more likely to follow that consensus versus defecting.

Second, the Green party (GPS) appears to have a distinct relationship for cause groups (NGOs) on defection. Similar to the FDP, Green MPs with more ties to cause groups are more likely to defect from their constituents on cross-pressured votes (for two types of referendum) and less on consensus votes; the sub-group effects for Green MPs are located outside of the main cluster of effects for most types of referendum. Again, as environmental interest groups likely fit into the category of "cause" groups, a similar story might explain the behavior seen by the Green MPs. When MPs

are more closely affiliated with groups corresponding to their party's primary stakeholder, they are more likely to follow the party when they are cross-pressured and toe the consensus position when one exists.

Finally, this provides suggestive evidence that the majority of the heterogeneous effects are driven by major parties. Noting that the SVP, CVP, SP, and FDP are the major parties in Switzerland (i.e. they constitute the government; Kriesi and Trechsel 2008), they comprise the majority of the distinct heterogeneous effects. Indeed, they constitute all-but-one of the heterogeneous economic effects. And, if the Green Party (GPS), the fifth-largest party, is included in the definition of major, five of the nine distinct effects for cause sub-groups come from a major party. I return to this point in the next section.

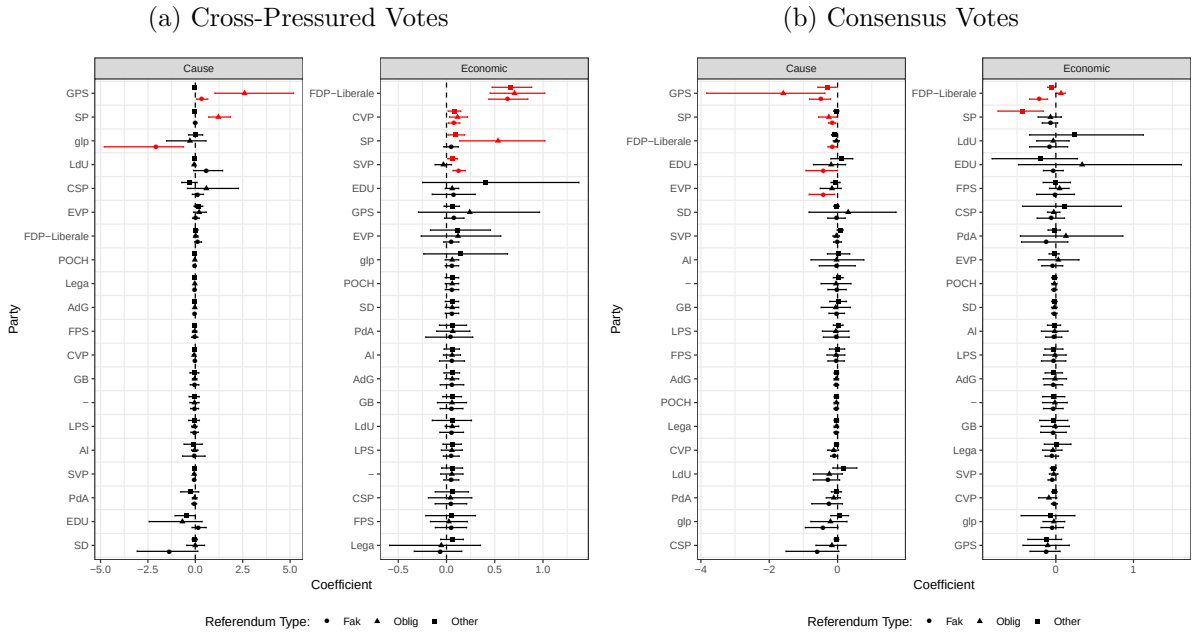
6.3 Examining Individual Heterogeneous Effects

Given that the groups are created by fusing together individual level units, it is also possible to examine the heterogeneous effects estimated for each party-and-referendum type. This is directly comparable to other existing methods that do not create groups. A key point to note that is these results will differ somewhat from the group-level analysis. That is because while a summary based on groups depends on all of the members and pools information across them, the individual analysis is focused on each unit. Thus, for units with more observations, they are likely to have smaller credible intervals as well as having different *means* even if the *mode* for the groups is shared. Applications of structured sparsity should consider both the groups that are formed as well as the unit-level effects.

Figure 3 does this by plotting the unit-level effects and the associated 95% credible interval; for cases where that does not include zero, they are shaded in red and put at the top of the figure. This individual-level analysis strengthens the group-level discussion above; we see that, as before, the FDP and Greens stand out for their distinctive patterns of significant effects. More interestingly, when the groups are disaggregated, a larger number of units that are put into the "Main Group" have a statistically significant effect. These tend to be the larger parties (e.g. SVP, CVP, and SP).

It is also interesting to note that these results give a relatively precise and small confidence interval to the effect for groups with little-to-no variation or with perfect separation. By contrast, the random effects approach (shown in Appendix F) gives a very wide credible interval for these

Figure 3: Heterogeneous Effects by Party and Referendum Type



Note: The 95% credible interval on each coefficient is shown with the posterior mean indicated by a marker. Groups for whom the interval does not contain zero are colored in red. Each panel shows the effect for cause and economic interest groups separately.

groups. The interval is much wider than the method estimated using structured sparsity; in the case of economic interest groups and cross-pressured votes, the narrow intervals for groups with little data range from -0.01 to 0.12 (in Figure 3a). By contrast, the same intervals span -0.75 to 0.65 for the random effects model. Thus, structured sparsity lets there be a more precise estimate of a limited effect for those groups versus a very uncertain estimate from traditional methods.

This difference comes from how information is shared; in the case of classical hierarchical models, pooling only occurs via the global shared random effect variance. By contrast, structured sparsity allows information to be directly shared across units that are connected (e.g. all units with the same referendum type). This, in some sense, is more plausible: If a new party were to emerge, it might be reasonable to expect that their effect is not simply the global effect plus a large mean-zero of noise, but rather can be more precisely estimated as a combination of their neighbors.

6.4 Implications from Examining Heterogeneous Effects

Broadly, however, these results identify heterogeneity that is centered around two dimensions; first, major parties seem to be driving the effects shown in the aggregate analysis. This is confirmed

by Table 6 which shows the results of sample splitting by major party. Second, the Green party appears to be distinct in terms of its effects—especially around cause groups. No clear patterns of heterogeneity emerge around referendum type. To examine these exploratory hypotheses more formally, I re-estimate the regression where I split the sample by major parties, the Green party, and minor parties to see whether the effects estimated by structure sparsity "hold up" when re-fitting the original model.

Table 6: Re-Analysis from Splitting the Data Based on Party Type

	Cross-Pressured Votes			*Consensus Votes*		
	Major Party	Green Party	Minor Party	Major Party	Green Party	Minor Party
Number of Economic Groups	0.127*	0.062	−0.044	−0.032*	−0.270	0.003
	(0.017)	(0.109)	(0.111)	(0.012)	(0.246)	(0.081)
Number of Cause Groups	−0.023	0.168*	−0.055	−0.043*	−0.521*	−0.157
	(0.020)	(0.076)	(0.080)	(0.020)	(0.149)	(0.093)
Observations	5,569	569	1,183	11,439	555	945

Note: Coefficients and standard errors reported. * indicates a p-value below 0.05. Models estimated using `glmer`. Models labelled with "major party" report results estimated on MPs from the SVP, CVP, SP, and FDP; "Green Party" is a model fit only on the Green Party and excludes a random effect for party. "Minor parties" include all other parties. Other controls are identical to that in Table 4 and outlined in Appendix F.

Table 6 shows the results and confirms that, as implied by the earlier analysis, major parties are driving the main effects—with only major parties and the Greens having any role of interest groups on defection for cross-pressured votes. Interestingly, on consensus votes, attachments to cause groups lead to *less* defection for all types of parties. This provides an important and substantively intelligible qualification to the results in Giger and Klüver (2016): Cause groups do lead to less defection but *only* in the case of consensus votes. That is, MPs with more cause group attachments are more likely to follow *both* their party and constituents when the chance presents itself—versus "going rogue". They do not, with the exception of the Green Party, have any statistically significant effect on encouraging less defection in the case where the MPs must make a hard decision.

7 Conclusion

This paper began by noting that, when modelling heterogeneity, it is often necessary to make simplifying assumptions to obtain relatively precise estimates of the sub-group effects of interest. A vast

array of methods provide different sets of assumptions to do this, mostly focused around stabilizing sub-group estimates around some global or aggregate effect. It argued that, while useful, a different assumption—stabilizing effects by creating data-driven groups guided by prior knowledge—may perform better. Existing work using structured sparsity provided a way to estimate these groups but it needed substantial modification to be suitable for social scientific research. To do this, I 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 some theoretical results about the propriety of the resulting posterior, 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, existing state-of-the-art methods struggled to accurately estimate the sub-group effects. Their approach of shrinking aggressively towards a global effect failed to do well insofar as that global effect did not well represent many groups. By contrast, structured sparsity—as well as traditional methods such as hierarchical models—performed well. This suggest that in cases where the researcher expects that underlying heterogeneity is based on groups, methods such as structured sparsity should be employed.

Second, I re-analyzed a recent paper (Giger and Klüver 2016) to explore whether there was heterogeneity in the effect of interest groups on MPs voting against their constituents. After re-structuring the data to better capture the underlying quantity of interest, I showed that there was clear evidence of heterogeneous effects by the MP's party and referendum type. Traditional methods of addressing heterogeneity (agnostic interactive models and random effects) fit less well than structured sparsity. Using this method to uncover heterogeneous effects showed two key points: First, major parties are driving most of the significant effects. It did not uncover much systematic evidence of effect heterogeneity by the type of referendum. Second, inside of those major parties, the groups uncovered at the structured sparse posterior mode clearly identified two parties (the FDP and the Greens) as having significantly more *homogeneity* in their MPs response to interest group affiliation. In substantive terms, the implications could be summarized as follows: When MPs are affiliated to interest groups corresponding to their primary stake-holder (economic groups for the major parties; cause groups for the Greens), they are more likely to choose their party's position over their constituents when the two conflict. While speculative, this would fit into the idea

of Green parties as being "distinct" from traditional parties in their closer ties to non-economic (vs. cause) groups. It would have been hard to discover this based on prior knowledge alone, especially the finding with respect to the FDP and Greens being distinct, without risking concerns about over-fitting based on noise; structured sparsity provides a principled way to identify these sub-groups with distinct effects.

Overall, this paper demonstrated that treating underlying heterogeneity as based on a small number of discrete groups can lead to benefits in estimating heterogeneous effects on both simulated and actual data. Whether the world is "truly" made up of discrete groups that share the unobserved heterogeneity is an open and somewhat philosophical question. All models of heterogeneity are abstractions from the truth and thus in some sense what matters fundamentally is whether they are useful to applied researchers. It is worth noting, however, that relying on discrete groups is fundamental to how researchers create concepts and make sense of the world; the deep incorporation of groups into our thinking may reflect something fundamental about the world that structured sparsity—or other methods based on simplification by creating groups—is able to uncover better than existing methods.

References

Albert, James H., and Siddhartha Chib. 1993. "Bayesian Analysis of Binary and Polychotomous Response Data". *Journal of the American Statistical Association* 88 (422): 669–679.

Ali, Alnur, and Ryan J. Tibshirani. 2019. "The Generalized Lasso Problem and Uniqueness". *Electronic Journal of Statistics* 13 (2): 2307–2347.

Andrews, D. F., and C. L. Mallows. 1974. "Scale Mixtures of Normal Distributions". *Journal of the Royal Statistical Society. Series B (Methodological)* 36 (1): 99–102.

Armagan, Artin, David B Dunson, and Jaeyong Lee. 2013. "Generalized double Pareto shrinkage". *Statistica Sinica* 23 (1): 119–143.

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.

Athey, Susan, and Guido Imbens. 2016. "Recursive Partitioning for Heterogeneous Causal Effects". *Proceedings of the National Academy of Sciences* 113 (27): 7353–7360.

Bach, Francis, et al. 2012. "Structured Sparsity through Convex Optimization". *Statistical Science* 27 (4): 450–468.

Barceló, Joan. 2017. "The Clarity of the Majority's Preference Moderates the Influence of Lobbying on Representation". *Party Politics* Advance Access:1–9. doi:10.1177/2F1354068817715803.

Bates, Douglas, et al. 2015. "Fitting Linear Mixed-Effects Models Using lme4". *Journal of Statistical Software* 67 (1): 1–48.

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.

Carey, John M. 2007. "Competing Principals, Political Institutions, and Party Unity in Legislative Voting". *American Journal of Political Science* 51 (1): 92–107.

Carey, John M., and Simon Hix. 2012. "District Magnitude and Representation of the Majority's Preferences: A Comment and Reinterpretation". *Public Choice* 154 (1-2): 139–148.

Carvalho, Carlos M., Nicholas G. Polson, and James G. Scott. 2010. "The Horseshoe Estimator for Sparse Signals". *Biometrika* 97 (2): 465–480.

Chen, Ming-Hui, and Qi-Man Shao. 2001. *Proceedings of the American Mathematical Society* 129 (1): 293–303.

Chen, Tian. 2015. "Computational Algorithms for Penalized Logistic Regression with Categorical Predictors and Random Effect Logistic Models". PhD thesis. `https://repository.lib.ncsu.edu/bitstream/handle/1840.16/10376/etd.pdf`.

Chen, Xi, et al. 2012. "Smoothing Proximal Gradient Method for General Structured Sparse Regression". *The Annals of Applied Statistics* 6 (2): 719–752.

Chernozhukov, Victor, et al. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters". *The Econometrics Journal* 21 (1): C1–C68.

Duan, Junbo, et al. 2016. "Generalized LASSO with Under-Determined Regularization Matrices". *Signal Processing* 127:239–246.

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.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent". *Journal of Statistical Software* 33 (1): 1–22.

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.

Gelman, Andrew, et al. 2008. "A Weakly Informative Default Prior Distribution for Logistic and Other regression models". *The Annals of Applied Statistics* 2 (4): 1360–1383.

Gertheiss, Jan, and Gerhard Tutz. 2010. "Sparse modeling of categorial explanatory variables". *The Annals of Applied Statistics* 4 (4): 2150–2180.

Giger, Nathalie, and Heike Klüver. 2016. "Voting Against Your Constituents? How Lobbying Affects Representation". *American Journal of Political Science* 60 (1): 190–205.

Green, Donald P., and Holger L. Kern. 2012. "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees". *Public Opinion Quarterly* 76 (3): 491–511.

Greene, William H., and Terry G. Seaks. 1991. "The Restricted Least Squares Estimator: A Pedagogical Note". *The Review of Economics and Statistics* 73 (3): 563.

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.

Hainmueller, Jens, and Chad Hazlett. 2014. "Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach". *Political Analysis* 22 (2): 143–168.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. 2nd. Springer-Verlang.

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright. 2015. *Statistical Learning with Sparsity*. Boca Raton: Chapman / Hall/CRC.

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.

Hui, Francis K. C., David I. Warton, and Scott D. Foster. 2015. "Tuning Parameter Selection for the Adaptive Lasso Using ERIC". *Journal of the American Statistical Association* 110 (509): 262–269.

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 2009*, 433–440.

Kriesi, Hanspeter, and Alexander H. Trechsel. 2008. *The Politics of Switzerland: Continuity and Change in a Consensus Democracy*. Cambridge University Press.

Kyung, Minjung, et al. 2010. "Penalized Regression, Standard Errors, and Bayesian Lassos". *Bayesian Analysis* 5 (2): 369–412.

Lawson, Charles L., and Richard J. Hanson. 1974. *Solving Least Squares Problems*. Englewood Cliffs, N.J.: Prentice-Hall.

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

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by randomForest". *R News* 2 (3): 18–22.

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

McLachlan, Geoffrey, and Thriyambakam Krishnan. 2008. *The EM Algorithm and Extensions*. 2nd. Wiley.

Meng, Xiao-Li, and David Van Dyk. 1997. "The EM Algorithm–an Old Folk-song Sung to a Fast New Tune". *Journal of the Royal Statistical Society. Series B (Methodological)* 59 (3): 511–567.

Meyer, David. 2019. *Support Vector Machines: The Interface to libsvm in package e1071*. Visited on. `https://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf`.

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.

Ninomiya, Yoshiyuki, and Shuichi Kawano. 2016. "AIC for the Lasso in generalized linear models". *Electronic Journal of Statistics* 10 (2): 2537–2560.

Oelker, Margret-Ruth, and Gerhard Tutz. 2017. "A uniform framework for the combination of penalties in generalized structured models". *Advances in Data Analysis and Classification* 11 (1): 97–120.

Park, Jaewoo, and Murali Haran. 2018. "Bayesian Inference in the Presence of Intractable Normalizing Functions". *Journal of the American Statistical Association* 113 (523): 1372–1390.

Park, Mee Young, and Trevor Hastie. 2007. "L1-regularization path algorithm for generalized linear models". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69 (4): 659–677.

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

Pauger, Daniela, and Helga Wagner. 2018. "Bayesian Effect Fusion for Categorical Predictors". *Bayesian Analysis* Advance Access. doi:`10.1214/18-ba1096`. `https://doi.org/10.1214/18-ba1096`.

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

Polson, Nicholas G., James G. Scott, and Jesse Windle. 2013. "Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables". *Journal of the American Statistical Association* 108 (504): 1339–1349.

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

Portmann, Marco, David Stadelmann, and Reiner Eichenberger. 2012. "District Magnitude and Representation of the Majority's Preferences: Evidence from Popular and Parliamentary Votes". *Public Choice* 153 (3–4): 585–610.

Ratkovic, Marc, and Dustin Tingley. 2018. "Causal Inference Through the Method of Direct Estimation". *Working Paper*. `https://scholar.harvard.edu/files/dtingley/files/mde-final.pdf`.

— . 2017. "Sparse Estimation and Uncertainty with Application to Subgroup Analysis". *Political Analysis* 25 (1): 1–40.

Roualdes, Edward. A. 2015. "Bayesian Trend Filtering". *arvix preprint*. `https://arxiv.org/pdf/1505.07710.pdf`.

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". Visited on. `https://shiraito.github.io/research/files/jmp.pdf`.

Simon, Noah, et al. 2013. "A Sparse-Group Lasso". *Journal of Computational and Graphical Statistics* 22 (2): 231–245.

Sparapani, Rodney, Charles Spanbauer, and Robert McCulloch. *The BART R package*. Visited on. `https://rdrr.io/cran/BART/f/inst/doc/the-BART-R-package.pdf`.

Speckman, Paul L., Jaeyong Lee, and Dongchu Sun. 2009. "Existence of the MLE and Propriety of Posteriors for a General Multinomial Choice Model". *Statistica Sinica* 19 (2): 731–748.

Stadelmann, David, Marco Portmann, and Reiner Eichenberger. 2013. "Quantifying parliamentary representation of constituents' preferences with quasi-experimental data". *Journal of Comparative Economics* 41 (1): 170–180.

Su, Liangjun, Zhentao Shi, and Peter C. B. Phillips. 2016. "Identifying Latent Structures in Panel Data". *Econometrica* 84 (6): 2215–2264.

Tansey, Wesley, et al. 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.

Tian, Lu, et al. 2014. "A Simple Method for Estimating Interactions Between a Treatment and a Large Number of Covariates". *Journal of the American Statistical Association* 109 (508): 1517–1532.

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

Tibshirani, Robert, et al. 2004. *The Annals of Statistics* 32 (2): 407–499.

Tibshirani, Ryan J., and Jonathan Taylor. 2012. "Degrees of freedom in lasso problems". *The Annals of Statistics* 40 (2): 1198–1232.

— . 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.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry. 2017. "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC". *Statistics and Computing* 27 (5): 1413–1432.

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.

Wang, Yu-Xiang, et al. 2016. "Trend Filtering on Graphs". *Journal of Machine Learning Research* 17 (105): 1–41.

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.

Zou, Hui. 2006. “The Adaptive LASSO and Its Oracle Properties”. 101 (476): 1418–1429.

Zou, Hui, Trevor Hastie, and Robert Tibshirani. 2007. “On the ”degrees of freedom” of the lasso”. *The Annals of Statistics* 35 (5): 2173–2192.

A Definition of Structured Sparsity

I use the term "structured sparsity" to refer to the methods outlined in this paper as it provides a conceptually clear way of articulating the core method and its relationship to traditional sparse models. Unfortunately, many different papers use different names to refer to similar models and thus it is important to exactly situate my use of "structured sparsity" alongside this research.

In the most general usage, I conceptualize structured sparsity as imposing sparsity on some number of linear and/or quadratic combinations of coefficients. If one uses the ℓ_1/ℓ_2 norm for penalization, this can be expressed as follows where $\boldsymbol{D} \in \mathbb{R}^{K \times p}$ and $\boldsymbol{F}_l$ is a real symmetric positive *semi*-definite matrix where $l \in \{1, \cdots, L\}$.

$$- ||\boldsymbol{D}\boldsymbol{\beta}||_1 - \sum_{l=1}^{L} ||\boldsymbol{\beta}^T \boldsymbol{F}_l \boldsymbol{\beta}||_2 \tag{9}$$

This usage is both slightly more general and more narrow than existing uses of "structured sparsity". For example, Bach et al. (2012) defines the penalty in terms of a ℓ_q norm (not necessarily an ℓ_2 norm) but also only focuses on penalizing groups of coefficients and thus many $\boldsymbol{D}$ would not fit into that framework. Chen et al. (2012) uses structured sparsity to focus on, separately, a specific choice of $||\boldsymbol{D}\boldsymbol{\beta}||$ and a specific choice of $\{\boldsymbol{F}_l\}$. Some work has combined both $\boldsymbol{D}$ and $\boldsymbol{F}_l$ penalties under names other than structured sparsity (e.g. "sparse group LASSO" Simon et al. 2013).

For the purposes of this paper, as I make explicit in main text, I focus only on the case of a penalty on linear combination of coefficients ($||\boldsymbol{D}\boldsymbol{\beta}||_1$). This corresponds to the penalty imposed by the generalized LASSO (Tibshirani and Taylor 2011). Unfortunately, there are again many different names for this penalty, especially for papers that consider particular choices of $\boldsymbol{D}$ (e.g. "graph LASSO" Jacob, Obozinski, and Vert 2009; "trend-filtering on a graph" Wang et al. 2016; and many others).

Although, I set aside detailed analysis of the combination of $\boldsymbol{D}$ and $\boldsymbol{F}_l$ penalties for future work, I note a few key points here. First, using the bounds in Lemma 1, the key theorems in the paper on propriety can likely be extended to cover the case of quadratic penalties ($\boldsymbol{F}_l$) with minimal difficulty. Second, sampling from the joint posterior remains tractable even if $\boldsymbol{F}$ are included by using results from Kyung et al. (2010). Finally, it is clear that the EM algorithm (Appendix C) can be easily extended to include L $\boldsymbol{F}_l$ matrices. Fleshing out the details of these extensions with

applications is an on-going area of research.

B Proof of Theorems on Structured Sparsity

This section proves the results in the main text about the properties of a structured sparse prior. To begin, I derive a number of lemmas that allow the main theorems to be proved.

B.1 Lemmas

Lemma 1 *For some $\boldsymbol{D} \in \mathbb{R}^{K \times p}$, the corresponding structured sparse prior is proper if and only if* $\text{rank}(\boldsymbol{D}) = p$ *(full column rank).*

I first note the following identity: $\boldsymbol{x} \in \mathbb{R}^k$, $||\boldsymbol{x}||_2 \leq ||\boldsymbol{x}||_1 \leq \sqrt{k}||\boldsymbol{x}||_2$. Using that result, the kernel of a structured sparse prior can be founded as follows:

$$\exp\left(-\lambda\sqrt{k}||\boldsymbol{D\beta}||_2\right) \leq \exp\left(-\lambda||\boldsymbol{D\beta}||_1\right) \leq \exp\left(-\lambda||\boldsymbol{D\beta}||_2\right) \tag{10}$$

Define $\boldsymbol{B} = \boldsymbol{D}^T\boldsymbol{D}$. Given that it is symmetric positive definite matrix, it can be decomposed using an eigen-decomposition where $\boldsymbol{Q}$ is some orthogonal matrix and $\boldsymbol{\Lambda}$ is a $\text{rank}(\boldsymbol{B})$ by $\text{rank}(\boldsymbol{B})$ diagonal matrix of the non-zero eigenvalues. Consider the integral of either bound with some $w > 0$.

$$\begin{aligned}
\int_{\mathbb{R}^p} \exp\left(-w||\boldsymbol{D\beta}||_2\right) d\boldsymbol{\beta} &= \int_{\mathbb{R}^p} \exp\left(-w\sqrt{\boldsymbol{\beta}^T\boldsymbol{D}^T\boldsymbol{D\beta}}\right) d\boldsymbol{\beta} && \text{(11a)}\\
&= \int_{\mathbb{R}^p} \exp\left(-w\sqrt{\boldsymbol{\beta}^T\boldsymbol{Q}^T\begin{bmatrix}\boldsymbol{\Lambda} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\boldsymbol{Q\beta}}\right) d\boldsymbol{\beta} && \text{(11b)}\\
&= \int_{\mathbb{R}^p} \exp\left(-w\sqrt{[\boldsymbol{\nu}_1, \boldsymbol{\nu}_2]\begin{bmatrix}\boldsymbol{\Lambda} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{\nu}_1\\ \boldsymbol{\nu}_2\end{bmatrix}}\right) d\boldsymbol{\nu}; \quad \boldsymbol{\nu} = \boldsymbol{Q\beta} && \text{(11c)}\\
&= \int_{\mathbb{R}^{\text{rank}(\boldsymbol{B})}} \exp\left(-w\sqrt{\boldsymbol{\nu}_1^T\boldsymbol{\Lambda\nu}_1}\right) d\boldsymbol{\nu}_1 \int_{\mathbb{R}^{p-\text{rank}(\boldsymbol{B})}} d\boldsymbol{\nu}_2 && \text{(11d)}
\end{aligned}$$

From this decomposition, it is clear that the integral diverges if $\text{rank}(\boldsymbol{B}) \neq p$. Since $\boldsymbol{B} = \boldsymbol{D}^T\boldsymbol{D}$, the integral only is finite if $\text{rank}(\boldsymbol{D}) = p$, i.e. it has full column rank. As the integral in Equation 11

upper and lower bounds the kernel of a structured sparse prior (with $w = \lambda$ and $w = \sqrt{k}\lambda$), $\text{rank}(\boldsymbol{D}) = p$ is a necessary and sufficient condition for the prior being proper. The lemma is thus established.

Lemma 2 *Define an offset structured sparse prior as follow: For any* $\boldsymbol{\mu} \in \mathbb{R}^k$,

$$p(\boldsymbol{\beta}) \propto \exp\left(-\lambda||\boldsymbol{D}\boldsymbol{\beta} + \boldsymbol{\mu}||_1\right) \tag{12}$$

The offset structured sparse prior is proper if and only if $\boldsymbol{D}$ *has full column rank.*

This result is proven as follows. First, necessity:[36] Assume that $\text{rank}(\boldsymbol{D}) \neq p$. Note that a singular value decomposition of $\boldsymbol{D}$ is employed in the first line and $\boldsymbol{x}_1$ indicates the first $\text{rank}(\boldsymbol{D})$ elements of $\boldsymbol{x}$.

$$\begin{aligned}
c &= \int_{\mathbb{R}^p} \exp\left(-\lambda||\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{\beta} + \boldsymbol{\mu}||_1\right) d\boldsymbol{\beta} && \text{(13a)}\\
&= \int_{\mathbb{R}^p} \exp\left(-\lambda||\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{x} + \boldsymbol{\mu}||_1\right) d\boldsymbol{x} \quad \boldsymbol{x} = \boldsymbol{V}^T\boldsymbol{\beta} && \text{(13b)}\\
&= \int_{\mathbb{R}^{\text{rank}(\boldsymbol{D})}} \exp\left(-\lambda||\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{x}_1 + \boldsymbol{\mu}||_1\right) d\boldsymbol{x}_1 \int_{\mathbb{R}^{p-\text{rank}(\boldsymbol{D})}} d\boldsymbol{x}_2 \quad \boldsymbol{x}^T = [\boldsymbol{x}_1, \boldsymbol{x}_2]^T && \text{(13c)}\\
& && \text{(13d)}
\end{aligned}$$

The integral clearly diverges with respect to $\boldsymbol{x}_2$ (i.e. $p - \text{rank}(\boldsymbol{D})$ elements) and thus the necessary condition is established.

Second, sufficiency can be proven as follows: Assume that $\boldsymbol{D} \in \mathbb{R}^{K \times p}$ is full column rank. Given that $\boldsymbol{D}$ is full column rank, it can be decomposed using a "thin singular value decomposition" such that $\boldsymbol{D} = \boldsymbol{U}_1\boldsymbol{\Sigma}\boldsymbol{V}^T$ where $\boldsymbol{U}_1$ consists of p orthonormal vectors, $\boldsymbol{\Sigma}$ is a full rank diagonal matrix, and $\boldsymbol{V}$ is an orthogonal matrix.

[36] For the ℓ_1 penalty, this can be shown simply by the triangle inequality:

$$\exp\left(-\lambda||\boldsymbol{D}\boldsymbol{\beta}||_1\right)\exp\left(-\lambda||\boldsymbol{\mu}||_1\right) \leq \exp\left(-\lambda||\boldsymbol{D}\boldsymbol{\beta} + \boldsymbol{\mu}||_1\right)$$

Clearly, if $\text{rank}(\boldsymbol{D}) \neq p$, the integral of the left-hand side is infinite by Lemma 1 and thus it is not proper.

$$c = \int_{\mathbb{R}^p} \exp\left(-\lambda||\boldsymbol{U}_1\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{\beta} + \boldsymbol{\mu}||_1\right) d\boldsymbol{\beta} \tag{14a}$$

$$= 1/det(\boldsymbol{\Sigma}) \int_{\mathbb{R}^p} \exp\left(-\lambda||\boldsymbol{U}_1\boldsymbol{x} + \boldsymbol{\mu}||_1\right) d\boldsymbol{x} \quad \boldsymbol{x} = \boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{\beta} \tag{14b}$$

$$= 1/det(\boldsymbol{\Sigma}) \int_{\boldsymbol{z}\in C(\boldsymbol{D})} \exp\left(-\lambda||\boldsymbol{z} + \boldsymbol{\mu}||_1\right) d\boldsymbol{z} \quad \boldsymbol{z} = \boldsymbol{U}_1\boldsymbol{x} \tag{14c}$$

$$\leq 1/det(\boldsymbol{\Sigma}) \int_{\boldsymbol{z}\in\mathbb{R}^K} \exp(-\lambda||\boldsymbol{z} + \boldsymbol{\mu}||_1) d\boldsymbol{z} \tag{14d}$$

$$= 1/det(\boldsymbol{\Sigma}) \int_{\boldsymbol{z}'\in\mathbb{R}^K} \exp(-\lambda||\boldsymbol{z}'||_1) d\boldsymbol{z}' \quad \boldsymbol{z}' = \boldsymbol{z} + \boldsymbol{\mu} \tag{14e}$$

$$= 1/det(\boldsymbol{\Sigma}) \int_{\boldsymbol{z}'\in\mathbb{R}^K} \exp\left(-\lambda \sum_{i=1}^{K} |z_i'|\right) d\boldsymbol{z}' < \infty \tag{14f}$$

Note that since $\boldsymbol{U}_1$ forms the basis for the column space of $\boldsymbol{D}$, Equation 14c simply represents Equation 14b in a different fashion. However, since the integrand is always positive, that integral is upper-bounded by the integral over the entire real space $\mathbb{R}^K$. From there, a final change of variables proves the result. The full rank of $\boldsymbol{D}$ is crucial to ensuring that $\boldsymbol{\Sigma}\boldsymbol{V}^T$ is a square invertible matrix and thus permitting the chang of variables to $\boldsymbol{x}$. Thus, if $\text{rank}(\boldsymbol{D}) = p$, the offset structured sparse prior is proper for any $\boldsymbol{\mu} \in \mathbb{R}^K$.

Thus, Lemma 2 is established.

Lemma 3 *For any $\boldsymbol{D} \in \mathbb{R}^{K\times p}$, the prior corresponding to structured sparsity can be decomposed into a **proper** structured sparse prior on a vector of length $\text{rank}(\boldsymbol{D}) \times 1$ and a flat (improper) prior on $p - \text{rank}(\boldsymbol{D})$ elements.*

This can be established using a singular-value decomposition (SVD) of $\boldsymbol{D}$.Recall that the SVD of any matrix $\boldsymbol{A}$ can be expressed as follows, where $\boldsymbol{U}$ is a $k \times k$ orthogonal matrix and $\boldsymbol{U}_1$ consists of the first $\text{rank}(\boldsymbol{A})$ columns. Similarly $\boldsymbol{V} = [\boldsymbol{V}_1, \boldsymbol{V}_2]$ is an orthogonal matrix and $\boldsymbol{V}_1$ consists of the first $\text{rank}(\boldsymbol{A})$ columns. $\boldsymbol{\Sigma}$ is a diagonal matrix of the $\text{rank}(\boldsymbol{A})$ non-zero singular values of $\boldsymbol{A}$.

$$\boldsymbol{A} = [\boldsymbol{U}_1\boldsymbol{U}_2] \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_1^T \\ \boldsymbol{V}_2^T \end{bmatrix} \tag{15}$$

To show the lemma, define the following orthogonal rotation of $\boldsymbol{\beta}$.

$$\tilde{\boldsymbol{\beta}} = \begin{bmatrix} \tilde{\boldsymbol{\beta}}_{\mathcal{C}} \\ \tilde{\boldsymbol{\beta}}_{\mathcal{N}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{V}_1^T \boldsymbol{\beta} \\ \boldsymbol{V}_2^T \boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} \boldsymbol{V}_1^T \\ \boldsymbol{V}_2^T \end{bmatrix} \boldsymbol{\beta} = \boldsymbol{V}^T \boldsymbol{\beta} \tag{16}$$

Note that for any real $\boldsymbol{\beta}$, the corresponding $\tilde{\boldsymbol{\beta}}_{\mathcal{N}}$ lies in the null space of $\boldsymbol{D}$. Similarly, $\tilde{\boldsymbol{\beta}}_{\mathcal{C}}$ lies in the column space of $\boldsymbol{D}^T$. Equation 17 shows the prior corresponding to this orthogonal rotation.

$$p(\tilde{\boldsymbol{\beta}}_{\mathcal{C}}, \tilde{\boldsymbol{\beta}}_{\mathcal{N}}) \propto \exp\left(-\lambda ||\tilde{\boldsymbol{D}}_{\mathcal{C}} \tilde{\boldsymbol{\beta}}_{\mathcal{C}}||_1\right); \quad \tilde{\boldsymbol{D}}_{\mathcal{C}} = \boldsymbol{U}_1 \boldsymbol{\Sigma} \tag{17}$$

Since $\boldsymbol{U}_1$ consists of rank($\boldsymbol{D}$) ortho-normal vectors (length $K \times 1$), it is full column rank. Since $\boldsymbol{\Sigma}$ is a rank($\boldsymbol{D}$) $\times$ rank($\boldsymbol{D}$) diagonal matrix, $\tilde{\boldsymbol{D}}_{\mathcal{C}}$ is full column rank. Thus, the prior is a proper prior on $\tilde{\boldsymbol{\beta}}_C$ (i.e. a rank($\boldsymbol{D}$)-length vector) and a flat prior on $\tilde{\boldsymbol{\beta}}_{\mathcal{N}}$ (i.e. a $p-$rank($\boldsymbol{D}$)-lenght vector). Thus the lemma is established.

Lemma 4 *Assume some vector space $\mathcal{V} \subseteq \mathbb{R}^n$ has dimension k and can be spanned by some basis $\mathcal{B}$. Two claims follows:*

- *For any vector $\boldsymbol{c} \in \mathbb{R}^k$, there is a unique vector $\boldsymbol{x} \in \mathcal{V}$ such that $\boldsymbol{x} = \mathcal{B}\boldsymbol{c}$*
- *For any vector $\boldsymbol{x} \in \mathcal{V}$, there is a unique vector $\boldsymbol{c} \in \mathbb{R}^k$ such that $\boldsymbol{x} = \mathcal{B}\boldsymbol{c}$*

The first follows automatically from noting that $\mathcal{B}\boldsymbol{c}$ produces a single vector. The second follows because $\mathcal{B}$ is a basis. It can be easily shown by contradiction: Assume that $\boldsymbol{c}$ and $\boldsymbol{c}'$, where $\boldsymbol{c} \neq \boldsymbol{c}'$, were such that $\mathcal{B}\boldsymbol{c} = \mathcal{B}\boldsymbol{c}' = \boldsymbol{x}$. This implies that $\mathcal{B}(\boldsymbol{c} - \boldsymbol{c}') = \boldsymbol{0}$. However, because $\mathcal{B}$ is made up of k linearly independent vectors, this can only hold if $\boldsymbol{c} = \boldsymbol{c}'$. This is a contradiction and thus the second result follows.

Lemma 5 *For any function $f : \mathbb{R}^n \to \mathbb{R}$. Assume that $\mathcal{B}$ is a basis for a k-dimensional vector space $\mathcal{V} \subseteq \mathbb{R}^n$, i.e. $\mathcal{B}$ consists of k linearly independent n-length vectors. Then, the following two claims are equivalent:*

1. *There is a unique and finite maximum of f at $\boldsymbol{x}^*$, i.e.*

$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x} \in \mathbb{R}^k} f(\boldsymbol{B}\boldsymbol{x}) \tag{18}$$

2. There is a unique and finite maximum of f amongst $\boldsymbol{y} \in \mathcal{V}$, call it $\boldsymbol{y}^*$

$$\boldsymbol{y}^* = \arg\max_{\boldsymbol{y} \in \mathcal{V}} f(\boldsymbol{y}) \tag{19}$$

This can be proven as follows. Consider when (1) holds: It is clear that $f(\boldsymbol{y}^*) = f(\boldsymbol{B}\boldsymbol{x}^*)$ if $\boldsymbol{y}^* = \boldsymbol{B}\boldsymbol{x}^*$ and thus $\boldsymbol{y}^*$ is one solution to Equation 19. Assume there was some other $\boldsymbol{y}' \neq \boldsymbol{y}^*$ such that $f(\boldsymbol{y}') \geq f(\boldsymbol{y}^*)$ (i.e. that (2) did not hold). By contradiction, this cannot be true: By Lemma 4, there is a unique $\boldsymbol{x}'$ that corresponds to $\boldsymbol{y}'$. By Equation 18, however, it must be the case that $f(\boldsymbol{x}') < f(\boldsymbol{x}^*)$ unless $\boldsymbol{x}' = \boldsymbol{x}^*$. Thus, a contradiction obtains and thus if (1) holds, then (2) holds.

Consider when (2) holds: By Lemma Three, $\boldsymbol{x}^*$ is the unique solution to $\boldsymbol{y}^* = \boldsymbol{B}\boldsymbol{x}^*$. As before, $\boldsymbol{x}^*$ is clearly a solution to Equation 18. Its uniqueness can again be argued by contradiction. Assume there was some $\boldsymbol{x}' \neq \boldsymbol{x}^*$ such that $f(\boldsymbol{B}\boldsymbol{x}') \geq f(\boldsymbol{B}\boldsymbol{x}^*)$. That implies that the corresponding $\boldsymbol{y}' = \boldsymbol{B}\boldsymbol{x}'$ is such that $f(\boldsymbol{y}') \geq f(\boldsymbol{y}^*)$. By (2) holding, however, this cannot occur as $\boldsymbol{y}^*$ is a unique maximum. Thus, $\boldsymbol{x}^*$ is the unique solution.

Lemma 6 *Assume that $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ and $\boldsymbol{D} \in \mathbb{R}^{m \times p}$. Define $\mathcal{B}_{\boldsymbol{A}}$ as a basis for the null space of some matrix $\boldsymbol{A}$.*

Define the set $\mathcal{S} \subseteq \mathbb{R}^p$ such that $\mathcal{S} = \{s : \boldsymbol{X}\boldsymbol{s} = \boldsymbol{0} \text{ and } \boldsymbol{D}\boldsymbol{s} = \boldsymbol{0}\}$. The following conditions are equivalent

1. *$\mathcal{S} = \{\boldsymbol{0}\}$, i.e. the only member of $\mathcal{S}$ is the zero vector $\boldsymbol{0}$.*
2. *$\mathcal{N}(\boldsymbol{X}) \cap \mathcal{N}(\boldsymbol{D}) = \{\boldsymbol{0}\}$*
3. *$\text{rank}\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{D} \end{pmatrix} = p$*
4. *$\text{rank}(\boldsymbol{X}\mathcal{B}_{\boldsymbol{D}}) = p - \text{rank}(\boldsymbol{D})$*
5. *$\text{rank}(\boldsymbol{D}\mathcal{B}_{\boldsymbol{X}}) = p - \text{rank}(\boldsymbol{X})$*

The proof follows by showing that they can all be restated as implying Condition (1).

1. By definition the null space of $\boldsymbol{A}$ is the set of vectors $\boldsymbol{c}$ such that $\boldsymbol{Ac} = \boldsymbol{0}$. Thus, for a vector to be a member of both the null space of $\boldsymbol{X}$ and $\boldsymbol{D}$, it must follow that: $\boldsymbol{Xc} = \boldsymbol{0}$ and $\boldsymbol{Dc} = \boldsymbol{0}$. This restates Condition (1).

2. The definition of full column rank states that the only solution to the following set of equations is the trivial one where $\boldsymbol{c} = \boldsymbol{0}$.

$$\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{D} \end{pmatrix} \boldsymbol{c} = \boldsymbol{0} \iff \boldsymbol{Xc} = \boldsymbol{0} \text{ and } \boldsymbol{Dc} = \boldsymbol{0}$$

   This is identical to Condition (1).

3. For $\boldsymbol{X}\mathcal{B}_{\boldsymbol{D}}$ to be full column rank, it must be the case that for $\boldsymbol{c}_{\boldsymbol{D}\in\mathbb{R}^{p-\text{rank}(\boldsymbol{D})}}$, the only solution to the following equation is the zero vector.

$$\boldsymbol{X}\mathcal{B}_{\boldsymbol{D}}\boldsymbol{c}_{\boldsymbol{D}} = \boldsymbol{0}$$

   I characterize the solutions to this equation. Define $\boldsymbol{a} = \mathcal{B}_{\boldsymbol{D}}\boldsymbol{c}_D$, i.e. $\boldsymbol{a} \in \mathcal{N}(\boldsymbol{D})$. For $\boldsymbol{Xa} = \boldsymbol{0}$, this requires, by definition, that $\boldsymbol{a} \in \mathcal{N}(\boldsymbol{X})$. Thus, the set of valid solutions are all points in the null space of $\boldsymbol{X}$ that are in the null space of $\boldsymbol{D}$. Thus, full rank requires that the null spaces only intersect in the trivial case of $\boldsymbol{0}$. This is identically Condition (2) and thus Condition (1).

4. $\boldsymbol{D}\mathcal{B}_{\boldsymbol{X}}$ can be shown to be equivalent to Condition (2) by an identical argument to that of Condition (4).

B.2 Proof of Theorems

B.2.1 Proof of Theorem 1

Theorem 1 can be proven straightforwardly from the above lemmas.First, the propriety if and only if $\text{rank}(\boldsymbol{D}) = p$ simply restates Lemma 1. Consider the case of a proper prior, i.e. $\text{rank}(\boldsymbol{D}) = p$. The normalizing constant can be determined by multiplying the kernel of the prior by $1/c$, defined in Equation 20:

$$c = \int \exp\left(-\lambda||\boldsymbol{D}\boldsymbol{\beta}||_1\right) d\boldsymbol{\beta} \tag{20a}$$

$$= \lambda^{-p} \int \exp\left(-||\boldsymbol{D}\boldsymbol{x}||_1\right) d\boldsymbol{x}; \quad \boldsymbol{x} = \lambda\boldsymbol{\beta} \tag{20b}$$

$$= \lambda^{-p} \cdot \frac{1}{w_{\boldsymbol{D}}} \tag{20c}$$

$$1/c = \lambda^p w_{\boldsymbol{D}} \quad 0 < w_{\boldsymbol{D}} < \infty \tag{20d}$$

The change of variables in 20b is appropriate because the integral is finite as c is finite because $\boldsymbol{D}$ is full column rank (Lemma 1). Thus, multiplying the kernel of a structured sparse prior by the reciprocal of the normalizing constant c establishes Theorem 1. The final remark about propriety if and only if $\text{rank}(\boldsymbol{D}) = p$ restates Lemma 1.

Consider the case where $\boldsymbol{D}$ is not full column rank, and the prior is improper. By Lemma 3 and the results above, the prior in the rotated space can be written as

$$p(\tilde{\boldsymbol{\beta}}_{\mathcal{C}}, \tilde{\boldsymbol{\beta}}_{\mathcal{N}}) \propto \lambda^{\text{rank}(\boldsymbol{D})} w_{\boldsymbol{D}} \exp\left(-\lambda||\tilde{\boldsymbol{D}}_{\mathcal{C}}\tilde{\boldsymbol{\beta}}_{\mathcal{C}}||_1\right); \quad \tilde{\boldsymbol{D}}_{\mathcal{C}} = \boldsymbol{U}_1\boldsymbol{\Sigma} \tag{21}$$

By transforming backwards, i.e. $\boldsymbol{V}_{\boldsymbol{D}}\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}$, this creates the form found in Theorem 1.

B.2.2 Proof of Theorem 2

Theorem 2 can be established by extending results from (Park and Casella 2008), see also (Kyung et al. 2010) and (Andrews and Mallows 1974). They note the following identity:

$$\int_0^\infty \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(\frac{-z^2}{2\tau^2} - \frac{\lambda^2\tau^2}{2}\right) \frac{\lambda^2}{2} d\tau^2 = \frac{\lambda}{2}\exp(-\lambda|z|) \tag{22a}$$

$$z \sim N(0, \tau^2); \quad \tau^2 \sim \text{Exp}(\lambda^2/2) \tag{22b}$$

This can be easily adapted to the case of structured sparsity as follows:

$$p(\boldsymbol{\beta}, \{\tau_k^2\}_{k=1}^K, \lambda) \propto w_{\boldsymbol{D}} \lambda^{rank(\boldsymbol{D})-K} \prod_{k=1}^K \frac{1}{\sqrt{2\pi\tau_k^2}} \exp\left(-\frac{\boldsymbol{\beta}^T\boldsymbol{d}_k\boldsymbol{d}_k^T\boldsymbol{\beta}}{2\tau_k^2} - \lambda^2/2 \cdot \tau_k^2\right) \lambda^2/2 \tag{23}$$

From this joint prior, the Gibbs Sampler in Theorem 2 follows immediately. Note, however, an important qualification. Assume that $\text{rank}(\boldsymbol{D}) = p$, i.e. the prior is proper. Then, the marginal prior on all $\{\tau_k^2\}_{k=1}^K$ is shown below, where $\boldsymbol{\tau}^{-1}$ stacks $1/\tau_k^2$ diagonally. Note that this prior is *not* usually made up of independent priors on each τ_k^2.

$$p\left(\{\tau_k^2\}_{k=1}^K|\lambda\right) \propto \det\left(2\pi \boldsymbol{D}^T\boldsymbol{\tau}^{-1}\boldsymbol{D}\right)^{-1/2} \prod_{k=1}^K \frac{\exp\left(-\lambda^2/2\cdot\tau_k^2\right)}{\sqrt{\tau_k^2}} \tag{24}$$

B.2.3 Proof of Theorem 3

The theorem assumes 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 $\boldsymbol{\beta} \in \mathbb{R}^p$.
- Prior: $p(\boldsymbol{\beta}) \propto \exp\left(-\lambda||\boldsymbol{D}\boldsymbol{\beta}||_1\right)$ where $\boldsymbol{D} \in \mathbb{R}^{K\times p}$

Consider the transformation discussed in Lemma 3, i.e. left-multiplying $\boldsymbol{\beta}$ by $\boldsymbol{V}$ coming from a singular value decomposition of $\boldsymbol{D}$—call this $\tilde{\boldsymbol{\beta}}$ as in Lemma 3. For clarity, I now refer to the rotation matrix as $\boldsymbol{V}_{\boldsymbol{D}}$. I denote the linear predictor that is a function of $\tilde{\boldsymbol{\beta}}$

$$p(\tilde{\boldsymbol{\beta}}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\tilde{\boldsymbol{\nu}}) \cdot \lambda^{\text{rank}(\boldsymbol{D})} w_{\boldsymbol{D}} \exp\left(-\lambda||\tilde{\boldsymbol{D}}_{\mathcal{C}}\tilde{\boldsymbol{\beta}}_{\mathcal{C}}||_1\right); \quad \tilde{\boldsymbol{\nu}} = \tilde{\boldsymbol{X}}_{\mathcal{C}}\tilde{\boldsymbol{\beta}}_{\mathcal{C}} + \tilde{\boldsymbol{X}}_{\mathcal{N}}\tilde{\boldsymbol{\beta}}_{\mathcal{N}} \tag{25}$$

Establishing the propriety of the posterior in Equation 25 can be done using results from Michalak and Morris (2016). The paper provides a number of critical results summarized below as the following lemma:

Lemma 7 *Michalak and Morris (2016): Assume a likelihood $L(\boldsymbol{\eta}|\boldsymbol{y})$ such that $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$. Define an exponentiated norm bound (ENB) as follows: An ENB hods if constants $c_0, c_1 > 0$ exist such that*[37]

$$L(\boldsymbol{\eta}|\boldsymbol{y}) \leq c_0 \exp\left(-c_1||\boldsymbol{\eta}||\right) \tag{26}$$

From this concept, Michalak and Morris (2016) derive and note a number of crucial results.

[37]They further remark (p. 550) that "the constants c_0 and c_1 can be chosen independently of the L_p norm, $p \geq 1$, because of norm equivalence, where two norms L_p and L_q on $\mathbb{R}^r$ are said to be norm-equivalent if and only if there exist constants $0 < c_2, c_2$ such that $c_2||\boldsymbol{v}||_p \leq ||\boldsymbol{v}||_q \leq c_3||\boldsymbol{v}_p$ for any vector $\boldsymbol{v}$. While c_0 and c_1 cannot depend on $\boldsymbol{\eta}$, they can depend on any know values including $\boldsymbol{y}$, $\boldsymbol{X}$, $\cdots$."

The relevant ones are summarized here. I restate them to align with the notation here and the specific models considered in this paper. The original results are referenced.

1. *If the likelihood as a function of $\boldsymbol{\eta}$ is log-concave and the MLE of $\boldsymbol{\eta}$ exists and is unique (or more broadly if $\boldsymbol{\eta}$ has multiple MLEs, all MLEs lie in a bounded set), then the likelihood has an ENB as a function of $\boldsymbol{\eta}$. (p. 550; Theorem 6, p. 561)*

2. *For fixed $\boldsymbol{y}$, assume a likelihood $L(\boldsymbol{\eta}|\boldsymbol{y})$ as defined above has an ENB as defined above. Assume that $\boldsymbol{X}$ is full column rank and the prior density on $\boldsymbol{\beta}$ is bounded, i.e. $p(\boldsymbol{\beta}) \leq M < \infty$. Then, the posterior distributions of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ are proper and $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ have proper posterior moment generating functions. (Theorem 1; p. 553).*

3. *If $\boldsymbol{\beta}$ is entirely or partially known, let $\boldsymbol{\beta} = [\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T]^T$ so that $\boldsymbol{\eta} - \boldsymbol{X}_2\boldsymbol{\beta}_2 = \boldsymbol{X}_1\boldsymbol{\beta}_1$ with $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$ partitioned accordingly with $\boldsymbol{\beta}_2$ known and $\boldsymbol{\beta}_1$ with Lebesque measure. This model has already been addressed by [the point above]. Because posterior propriety holds for all fixed $\boldsymbol{\beta}_2$, it holds when $\boldsymbol{\beta}_2$ has a proper prior distribution. (Remark 9; p. 559).*

4. *GLMs [generalized linear models] with natural links are log-concave. Thus, a likelihood function $L(\boldsymbol{\eta}|\boldsymbol{y})$ for a GLM with a natural link and a finite MLE has an ENB as a function of $\boldsymbol{\eta}$. More generally, given a GLMM [generalized linear mixed model] (or other model) with a log-concave likelihood, it has an ENB if the MLE of $\boldsymbol{\eta}$ exists. (p. 551)*

This lemma is crucial to establishing posterior propriety. As it is phrased in a rather general way, some remarks are in order to make clear the relevance to this paper. First, for the models considered, I assume that we are focused on models where the likelihood is log-concave. This includes most generalized linear models with standard choices of link functions. In general, their results could be applied to more complex models, but I leverage Remarks (1) and (4) from Lemma 7 to use the existence of the MLE of $\boldsymbol{\eta}$ (and corresponding existence and uniqueness) to derive simple, easily verifiable, sufficient conditions for posterior propriety. Structured sparsity could be applied to more general models, but this requires more work to establish clear conditions for assessing the existence of an ENB.

Second, note that the results are stated in terms of the existence and uniqueness of the MLE on $\boldsymbol{\eta}$ *not* $\boldsymbol{\beta}$. This is designed to deal with the case of a rank deficient $\boldsymbol{X}$; adapting an example from

Michalak and Morris (2016, p. 552), imagine that $\boldsymbol{X}$ had two identical columns. The MLE of $\boldsymbol{\beta}$ is clearly not unique although the MLE of $\boldsymbol{\eta}$ could be—as any MLE leads to the same $\boldsymbol{X\beta}$.

Returning to the structured sparse case, note that Equation 25 reflects the scenario described in Remark (3) of Lemma 7 where $\tilde{\boldsymbol{\beta}}_{\mathcal{C}}$ has a proper prior and there is a flat prior on $\tilde{\boldsymbol{\beta}}_{\mathcal{N}}$. I prove the following Lemma:

Lemma 8 *For the likelihood on $\tilde{\boldsymbol{\nu}}$ described in Equation 25, define $\hat{\tilde{\boldsymbol{\beta}}}_{\mathcal{N}}$ as follows.*

$$\hat{\tilde{\boldsymbol{\beta}}}_{\mathcal{N}} = \arg\max_{\boldsymbol{\gamma}} \ln f(\boldsymbol{y}|\boldsymbol{\psi}); \quad \boldsymbol{\psi} = \tilde{\boldsymbol{X}}_{\mathcal{N}}\boldsymbol{\gamma} \equiv \boldsymbol{X}\boldsymbol{V}_{\boldsymbol{D},\mathcal{N}}\boldsymbol{\gamma} \tag{27}$$

If $\hat{\tilde{\boldsymbol{\beta}}}_{\mathcal{N}}$ exists and is unique, then the posterior on $\tilde{\boldsymbol{\beta}}$ (and thus $\boldsymbol{\beta}$) is proper.

This can be proved by directly applying Michalak and Morris (2016)'s results summarized in Lemma 7. If the MLE on $\tilde{\boldsymbol{\beta}}_{\mathcal{N}}$ exists and is unique, then this ensures that the MLE on $\tilde{\boldsymbol{\nu}}$ exists and is unique. Thus, Remarks (1) and (2) from Lemma 7 apply and the posterior is proper for $\tilde{\boldsymbol{\beta}}_{\mathcal{C}} = \boldsymbol{0}$. Note, however, that for any choice of $\tilde{\boldsymbol{\beta}}_{\mathcal{C}} \in \mathbb{R}^{\text{rank}(\boldsymbol{D})}$, the posterior remains proper as if the MLE of $\tilde{\boldsymbol{\nu}}$ exists and is unique when $\tilde{\boldsymbol{\beta}}_{\mathcal{C}} = \boldsymbol{0}$, it can be simply shifted to account for a non-zero offset.[38] Thus, Remark (3) from Lemma 7 applies as there is a proper prior on $\tilde{\boldsymbol{\beta}}_{\mathcal{C}}$ (by Lemma 1) and thus the posterior on $\tilde{\boldsymbol{\beta}}$ and thus $\boldsymbol{\beta}$ is proper.

Condition (b) in Theorem 3 expresses the claim in Lemma 8 slightly differently. It states that if $\hat{\boldsymbol{\beta}}_{\mathcal{N}(\boldsymbol{D})}$, as defined below, exists and is unique, then the posterior is proper.

$$\hat{\boldsymbol{\beta}}_{\mathcal{N}(\boldsymbol{D})} = \arg\max_{\boldsymbol{\beta} \in \mathcal{N}(\boldsymbol{D})} L(\boldsymbol{\eta}|\boldsymbol{y}); \quad \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} \tag{28}$$

The equivalence between this condition and the one defined in Lemma 8 follows by Lemma 5 as $\boldsymbol{V}_{\boldsymbol{D},\mathcal{V}}$ is a basis for the null space of $\boldsymbol{D}$. Thus, verifying that Lemma 8 holds is equivalent to checking that $\hat{\tilde{\boldsymbol{\beta}}}_{\mathcal{N}(\boldsymbol{D})}$ exists and is unique. Thus, (b) is sufficient for posterior propriety.

A necessary condition can be developed by examining the rank of $\tilde{\boldsymbol{X}}_{\mathcal{N}}$. This must be full rank for the posterior to be proper; the proof proceeds by contradiction: Assume that $\text{rank}(\tilde{\boldsymbol{X}}_{\mathcal{N}}) < p - \text{rank}(\boldsymbol{D})$, i.e. it was not full column rank. Posterior propriety turns on the integrability of

[38] Put another way, if the MLE exists and is unique, then any finite "offset" will still have an MLE that exists and is unique.

Equation 25 w.r.t. all of its arguments:

$$c = \int f(\boldsymbol{y}|\tilde{\boldsymbol{\eta}})\lambda^{\text{rank}(\boldsymbol{D})} w_{\boldsymbol{D}} \exp\left(-\lambda||\tilde{\boldsymbol{D}}_{\mathcal{C}}\tilde{\boldsymbol{\beta}}_{\mathcal{C}}||_1\right) d\tilde{\boldsymbol{\beta}}_{\mathcal{C}} d\tilde{\boldsymbol{\beta}}_{\mathcal{N}}; \quad \tilde{\boldsymbol{\nu}} = \tilde{\boldsymbol{X}}_{\mathcal{C}}\tilde{\boldsymbol{\beta}}_{\mathcal{C}} + \tilde{\boldsymbol{X}}_{\mathcal{N}}\tilde{\boldsymbol{\beta}}_{\mathcal{N}} \tag{29a}$$

$$= \int f(\boldsymbol{y}|\tilde{\boldsymbol{\eta}}')\lambda^{\text{rank}(\boldsymbol{D})} w_{\boldsymbol{D}} \exp\left(-\lambda||\tilde{\boldsymbol{D}}_{\mathcal{C}}\tilde{\boldsymbol{\beta}}_{\mathcal{C}}||_1\right) d\tilde{\boldsymbol{\beta}}_{\mathcal{C}} d\tilde{\tilde{\boldsymbol{\beta}}}_{\mathcal{N}} d\boldsymbol{\xi}; \quad \tilde{\boldsymbol{\nu}}' = \tilde{\boldsymbol{X}}_{\mathcal{C}}\tilde{\boldsymbol{\beta}}_{\mathcal{C}} + \tilde{\tilde{\boldsymbol{X}}}_{\mathcal{N}}\tilde{\tilde{\boldsymbol{\beta}}}_{\mathcal{N}} \tag{29b}$$

$$= \int_{\mathbb{R}^{p-m_\xi}} f(\boldsymbol{y}|\tilde{\boldsymbol{\eta}}')\lambda^{\text{rank}(\boldsymbol{D})} w_{\boldsymbol{D}} \exp\left(-\lambda||\tilde{\boldsymbol{D}}_{\mathcal{C}}\tilde{\boldsymbol{\beta}}_{\mathcal{C}}||_1\right) d\tilde{\boldsymbol{\beta}}_{\mathcal{C}} d\tilde{\tilde{\boldsymbol{\beta}}}_{\mathcal{N}} \times \int_{\mathbb{R}^{m_\xi}} 1 \cdot d\boldsymbol{\xi} \tag{29c}$$

$$m_\xi = p - \text{rank}(\boldsymbol{D}) - \text{rank}(\tilde{\boldsymbol{X}}_{\mathcal{N}})$$

Equation 29 follows from a change of variables by multiplying $\tilde{\boldsymbol{\beta}}_{\mathcal{N}}$ by an orthogonal matrix $\boldsymbol{V}'$ from an SVD of $\tilde{\boldsymbol{X}}_{\mathcal{N}}$. Since $\text{rank}(\tilde{\boldsymbol{X}}_{\mathcal{N}}) < p - \text{rank}(\boldsymbol{D})$, there are $m_\xi = p - \text{rank}(\boldsymbol{D}) - \text{rank}(\tilde{\boldsymbol{X}}_{\mathcal{N}}) > 0$ arguments that do not appear in the integrand. As the integral is over the real line, it diverges w.r.t. those elements and thus the posterior is not proper. Thus, integrability and posterior propriety requires $\tilde{\boldsymbol{X}}_{\mathcal{N}}$ to be full column rank.

This is thus a necessary condition for posterior propriety. By Lemma 6, $\text{rank}(\tilde{\boldsymbol{X}}_{\mathcal{N}})$ is equivalent to condition (a) in Theorem 3, e.g. $\mathcal{N}(\boldsymbol{X}) \cap \mathcal{N}(\boldsymbol{D}) = \{\boldsymbol{0}\}$. Thus, the necessity of (a) is established.

Note that this demonstrates why existence of $\hat{\boldsymbol{\beta}}_{\mathcal{N}(\boldsymbol{D})}$ is **not** sufficient on its own. Imagine there were multiple MLEs. Because of the log-concavity, this occurs because of a rank deficiency in $\tilde{\boldsymbol{X}}_{\mathcal{N}}$. That means, however, that a necessary condition for posterior propriety is violated and thus the posterior is improper. Thus, one could restate Theorem 2 as follows: condition (b′) requires the existence of $\hat{\boldsymbol{\beta}}_{\mathcal{N}(\boldsymbol{D})}$ and then (b′) and (a) are jointly sufficient for posterior propriety.

B.2.4 Proof of Corollary 1

Corollary 1 can be proven straightforwardly. For the linear model, assume Condition (a) holds. This implies that $\tilde{\boldsymbol{X}}_{\mathcal{N}}$ is full rank by Lemma 6. For a linear model, that ensures a single unique MLE and thus (a) implies (b). As (b) is sufficient for propriety, (a) is necessary and sufficient for posterior propriety. (a) and (b) is thus a slightly redundant way of stating this claim.

For the multinomial case with a standard link (logit or probit), results in Speckman, Lee, and Sun (2009) can be employed. Specifically, Theorem 3 (p. 742) states that

> For the multinomial logistic or probit choice model, the following conditions are equi-

valent.

1. There is overlap in the sample [see paper for discussion]
2. The MLE of $\boldsymbol{\beta}$ exists and is finite.
3. The posterior of $\boldsymbol{\beta}$ is proper under the constant prior

Assume that Conditions (a) and (b) hold. Again, $\tilde{\boldsymbol{X}}_{\mathcal{N}}$ is full rank. Thus, if the MLE exists, it is unique. Thus, this implies that the posterior on $\tilde{\boldsymbol{\beta}}_{\mathcal{N}}$ is proper. Propriety of the full posterior on $\boldsymbol{\beta}$ is ensured by the assumption of Condition (a). Thus, (a) and (b) are jointly sufficient for posterior propriety; they are jointly necessary because (a) is necessary. For binary models with other links, results in papers such as Chen and Shao (2001) can be employed.

B.2.5 Proof of Theorem 4

Recall that Theorem 4 made the following claims for a global-local prior:

- The prior on $\boldsymbol{\beta}$ is proper if and only if $\text{rank}(\boldsymbol{D}) = p$.
- Theorem 2 characterizes posterior propriety for any permissible g.
- If proper, the prior can be sampled where $\{\tau_k^2\}_{k=1}^K|\boldsymbol{\beta}$ is in the same family as in the standard sparse case (i.e.$\boldsymbol{D} = \boldsymbol{I}$) and $\boldsymbol{\beta}|\{\tau_k^2\}_{k=1}^K$ is normally distributed.

First, it is worth noting that this joint prior induces sparsity on $\boldsymbol{d}_k^T\boldsymbol{\beta}$, i.e. structured sparsity. This follows automatically by a slight re-arrangement of the joint prior. Recall that a global-local prior implies some density on δ, i.e. $p_{g,\boldsymbol{\lambda}}(\delta)$, such that the posterior mode encourages $\delta = 0$. The subscripts $g, \boldsymbol{\lambda}$ note the dependence of this marginal prior on the mixing distribution g and the hyper-parameters $\boldsymbol{\lambda}$. Calculating the marginal (possibly improper) prior on $\boldsymbol{\beta}$ can be found by integrating away the τ^2 and yielding the following result:

$$p(\boldsymbol{\beta}) \propto \prod_{k=1}^{K} p_{g,\boldsymbol{\lambda}}(\boldsymbol{d}_k^T\boldsymbol{\beta}) \tag{30}$$

Thus, it is clear that the posterior mode with this prior encourages $\boldsymbol{d}_k^T\boldsymbol{\beta} = 0$ and thus implies structured sparsity. Note, however, that this does *not* say that the prior for global-local structured

sparsity is *identical* to the product of K i.i.d. priors for global-local sparsity. Rather, there is a complicated normalizing constant that depends on $\boldsymbol{\lambda}, \boldsymbol{D}$ and g. Thus, inference on the hyper-parameters may be more complicated and needs further work to understand outside of the LASSO case.

Second, the first two claims (propriety of prior and posterior) follow by noting that Lemma 2 with $\boldsymbol{\mu} = \mathbf{0}$ and Theorem 2 does not use anything besides the structure of $\boldsymbol{D}$ to establish finiteness and thus applies automatically to global-local priors generally.

C Details of Inference

This section derives the Gibbs Sampler algorithms for the linear and multinomial models. It then discusses particularities of the EM algorithm.

C.1 Linear Regression

For linear regression, we need to incorporate the error variance σ^2 into the model. Assume the following generative framework following Park and Casella (2008):

$$\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2 \sim N\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_N\right) \tag{31a}$$

$$p(\boldsymbol{\beta}|\lambda^2, \sigma^2) \propto w_{\boldsymbol{D}} \lambda^m / \sigma^m \exp\left(-\frac{\lambda}{\sigma}||\boldsymbol{D}\boldsymbol{\beta}||_1\right) \tag{31b}$$

The log-posterior, including a prior of $p_0(\sigma^2)$ on σ^2 and $p_0(\lambda^2)$ on λ^2 can be written as follows, up to constant involving $\boldsymbol{D}$:

$$-\frac{N}{2}\ln(2\pi\sigma^2) - \frac{||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2}{2\sigma^2} + m\ln(\lambda) - m/2\ln(\sigma^2) - \frac{\lambda}{\sigma}||\boldsymbol{D}\boldsymbol{\beta}||_1 + \ln p_0(\sigma^2) + \ln p_0(\lambda^2) \tag{32}$$

With data augmentation, we can write the prior on $\boldsymbol{\beta}, \boldsymbol{\tau}|\sigma^2$ as follows; this follows the advice in Park and Casella (2008) to have the prior depend on σ^2.[39]

[39]The fact that σ^2 is to the $-m/2$ power can be see more immediately if one looks at the rotated $\boldsymbol{\beta}$ space where the normal is non-singular.

$$p(\boldsymbol{\beta}, \boldsymbol{\tau}|\lambda, \sigma^2) \propto (\sigma^2)^{-m/2}\lambda^{K+m}det(\boldsymbol{\tau})^{-1/2}\exp(\sum_k -\lambda^2/2\tau_k^2)\exp\left(-\frac{\boldsymbol{\beta}^T\boldsymbol{D}^T\boldsymbol{\tau}^{-1}\boldsymbol{D}\boldsymbol{\beta}}{2\sigma^2}\right) \tag{33}$$

From this, the full augmented log-posterior can be written as:

$$-\frac{N}{2}\ln(2\pi\sigma^2) - \frac{||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}||_2^2}{2\sigma^2} + (K+m)/2\ln(\lambda^2) - \sum_k \tau_k^2\lambda^2/2 - 1/2\sum_k \ln(\tau_k^2) \tag{34a}$$

$$-m/2\ln(\sigma^2) - \frac{\boldsymbol{\beta}^T\boldsymbol{D}^T\boldsymbol{\tau}^{-1}\boldsymbol{D}\boldsymbol{\beta}}{2\sigma^2} \tag{34b}$$

From this, we see that the full conditional for σ^2 becomes, assuming a conjugate prior of $p_0(\sigma^2) \sim \text{InverseGamma}(a_0, b_0)$:

$$\sigma^2|- \sim \text{InverseGamma}\left(a_{0,\sigma} + \frac{1}{2}\left[N + Rank(\boldsymbol{D})\right], b_{0,\sigma} + \frac{1}{2}\left[(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{\beta}^T\boldsymbol{D}^T\boldsymbol{\tau}^{-1}\boldsymbol{D}\boldsymbol{\beta}\right]\right) \tag{35}$$

The full conditionals on the other parameters are easily derived, where $\boldsymbol{\tau}$ is a $K \times K$ diagonal matrix with each element being τ_k^2.

$$\boldsymbol{\beta}|. \sim N\left(\boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{y}, \sigma^2\boldsymbol{A}^{-1}\right); \quad \boldsymbol{A} = (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\tau}^{-1}) \tag{36a}$$

$$1/\tau_k^2 \sim \text{InvGaussian}\left(\frac{\lambda\sigma}{|\boldsymbol{d}_k^T\boldsymbol{\beta}_j|}, \lambda^2\right) \tag{36b}$$

$$\lambda^2 \sim \text{Gamma}\left(a_{0,\Lambda} + [K+m]/2, \quad b_{0,\Lambda} + \frac{1}{2}\sum_{k=1}^K \tau_k^2\right) \tag{36c}$$

C.2 Multinomial Regression

Inference is derived for a C-category multinomial regression with the logistic regression being a special case. Denote the observation as y_i as taking on values from 1 to C. For simplicity, I assume the covariates are equal across levels. For each y_i, the generative model is multinomial:

$$p(y_i = c|\{\boldsymbol{\beta}_c\}) \propto \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}_c) \tag{37}$$

The likelihood is shown below, setting $\boldsymbol{\beta}_C = \mathbf{0}$ to identify the model.

$$\prod_{i=1}^{N} \left[\frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_c)}{\sum_{l=1}^{C} \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_c)} \right]^{I(y_i=c)} \tag{38}$$

Structured sparsity, as before, can be encoded by placing priors on $\boldsymbol{\beta}_c$. I focus on the case of identical $\boldsymbol{D}$ for each $\boldsymbol{\beta}_c$, but one could impose more complex restrictions by constraining coefficients across-levels c but care would need to be taken to ensure that the baseline level c does not matter.

Assume, therefore, that the prior structure has the form

$$p(\{\boldsymbol{\beta}_c\}) \propto \prod_{c=1}^{C-1} \lambda^m \exp(-\lambda ||\boldsymbol{D}\boldsymbol{\beta}_c||) \tag{39}$$

A Gibbs Sampler can be constructed following Polson, Scott, and Windle (2013).[40] The intuition is that, for each c (except the baseline), we want to do inference on the following posterior:

$$p(\{\boldsymbol{\beta}_c\}|\{\boldsymbol{\beta}_{\neg c}\}) = \prod_{i=1}^{N} \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_c - O_{ic})^{I(y_i=c)}}{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_c - O_{ic}) + 1} \cdot p(\{\boldsymbol{\beta}_c\}_{t=1}^{T}) \tag{40}$$

$$O_{ic} = \ln \left(\sum_{l \neq c} \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_l) \right)$$

From this, one can do Polya-Gamma augmentation as outlined in Polson, Scott, and Windle (2013). The core identity is that, for $\omega \sim PG(1, x)$ where PG is a Polya-Gamma random variable—a particular infinite convolution of Gamman random variables:[41]

$$\boldsymbol{\beta}_c|\{\boldsymbol{\beta}_{\neg c}\} = \prod_{i=1}^{N} \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_c - C_{ic})^{I(y_i=c)}}{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_c - O_{ic}) + 1} \cdot p(\{\boldsymbol{\beta}_{c'}\}_{c'=1}^{C}) \tag{41a}$$

$$\omega_{i,c}|\boldsymbol{\beta}_c, \{\boldsymbol{\beta}_{\neg c}\} \sim PG(1, \boldsymbol{x}_i^T \boldsymbol{\beta}_c - O_{ic}) \tag{41b}$$

$$\boldsymbol{\beta}_c|\{\omega_{i,c}\}, \{\boldsymbol{\beta}_{\neg c}\} \sim N\left(\boldsymbol{\Lambda}_\beta^{-1} \boldsymbol{X}^T \boldsymbol{s}, \quad \boldsymbol{\Lambda}_\beta^{-1} \right) \tag{41c}$$

[40] The EM algorithm is more involved here but still has convergence properties as it is an AECM (Meng and Van Dyk 1997). [Working paper by author] provides a detailed derivation.

[41] Specifically, see Polson, Scott, and Windle (2013) for details.

$$\omega = \frac{1}{2\pi^2} \sum_{i=1}^{\infty} \frac{Z_i}{(k-1/2)^2 + c^2/(4\pi^2)}; \quad Z_i \sim^{i.i.d.} Gamma(b, 1)$$

$$\boldsymbol{\Lambda}_\beta = \left[\sum_i \omega_{i,c} \boldsymbol{x}_i \boldsymbol{x}_i^T\right]$$

$$s_i = I(y_i = c) - 1/2 - \omega_i(\boldsymbol{x}_i^T \boldsymbol{\beta}_c - O_{ic}); \quad [\boldsymbol{s}]_i = s_i$$

The key point of this manipulation is that the data augmentation for the outcome, i.e. to make it conditionally normal given the Polya-Gamma augmentation variables, occurs independently of the data augmentation for the sparsity penalty. Thus, one can sample the $\{\tau_k^2\}$ as before and thus create a posterior on $\boldsymbol{\beta}_c$ as follows

$$\boldsymbol{\beta}_c | \{\omega_{i,c}\}, \{\boldsymbol{\beta}_{\neg c}\}, \{\tau_k^2\} \sim N\left(\left[\boldsymbol{\Lambda}_\beta + \boldsymbol{D}^T \boldsymbol{\tau}^{-1} \boldsymbol{D}\right]^{-1} \boldsymbol{X}^T \boldsymbol{s}, \quad \left[\boldsymbol{\Lambda}_\beta + \boldsymbol{D}^T \boldsymbol{\tau}^{-1} \boldsymbol{D}\right]^{-1}\right) \tag{42}$$

C.3 EM Algorithm

The EM algorithm cycles through the full conditionals in the following way. First, it takes the log-full conditional for $\boldsymbol{\beta}$ given $\{\tau_k^2\}$ and $\{\omega_{i,c}\}$ in the case of the multinomial model; it then plugs in the relevant expectations of τ_k^2 and $\{\omega_{i,c}\}$ using their own full conditional distributions. This is the E-Step. The key point is to note that for an Inverse-Gaussian, the moments are as follow if $1/\tau_k^2 \sim \text{InverseGaussian}(\mu, \lambda)$.

$$E[1/\tau_k^2] = \mu; \quad E[\tau_k^2] = 1/\lambda + 1/\mu \tag{43}$$

In the case of structured sparsity, $E[1/\tau_k^2] = \frac{1}{|\boldsymbol{d}_k^T \boldsymbol{\beta}|}$. After the E-Step, the algorithm maximizes the log-full conditional posterior w.r.t. $\boldsymbol{\beta}$; this is a simple least-squares (ridge) update. This is the M-Step. It iterates these until either log-posterior increases by a small amount or the estimated $\boldsymbol{\beta}$ does not change noticeably.

As the EM algorithm suggests, if $\boldsymbol{d}_k^T \boldsymbol{\beta} = 0$ or $\boldsymbol{\beta}^T \boldsymbol{F} \boldsymbol{\beta} = 0$, then the corresponding augmentation variable τ_k^2 no longer has a proper density. For the fully Bayesian approach, this will not arise (as $\boldsymbol{d}_k^T \boldsymbol{\beta}$ will never *exactly* equal zero), but this will occur in the EM algorithm. As Polson and Scott (2011b) note in the standard sparse case, this is a feature not a bug—$\boldsymbol{d}_k^T \boldsymbol{\beta} = 0$ implies that the restriction holds at the posterior mode and the variables are sparsified as implied by $\boldsymbol{d}_k$. For inference, however, one must deal with the fact that as $\boldsymbol{d}_k^T \boldsymbol{\beta} \to 0$, $E[1/\tau_k^2] \to \infty$. There are two strategies available in the associated software. Both approaches assume that if $|\boldsymbol{d}_k^T \boldsymbol{\beta}| < \epsilon$ where ϵ

is some small positive constant, it should be treated as "infinite" and dealt with appropriately.

1. Restricted Least Squares: This option modifies the approach in Polson and Scott (2011b). They suggest that, in the standard case, once $|\beta_j| < \epsilon$, then one should run restricted least squares (e.g. Greene and Seaks 1991), i.e. OLS where $\beta_j = 0$ must hold. In their case, this can be solved by an expanded least squares system derived from Lagrange multipliers. The analogue for structured sparsity is that, if $|\boldsymbol{d}_k^T\boldsymbol{\beta}| < \epsilon$, then require those restrictions to hold. Define $\boldsymbol{D}_{\neg\mathcal{A}_n}$ as the rows of $\boldsymbol{D}$ where this holds. The rank of $\boldsymbol{D}_{\neg\mathcal{A}_n}$ is $m_{\neg\mathcal{A}}$.

   The slight complication, however, is that as Greene and Seaks (1991) notes, their procedure works only if $\boldsymbol{D}_{\neg\mathcal{A}_n}$ has linearly independent rows. This is always true in the standard case that Polson and Scott (2011b) consider, see the discussion elsewhere in the paper, but will not occur for many important cases (including the ones analyzed in the empirical section!). Thus, some additional work is needed.

   Broadly, there are two strategies. The first point to note that is $\boldsymbol{\beta}$ must lie in the null space of $\boldsymbol{D}_{\neg\mathcal{A}_n}$ by definition. The first strategy targets this directly (Lawson and Hanson 1974); we can characterize the null space in terms of, say, a singular value decomposition of $\boldsymbol{D}_{\neg\mathcal{A}_n}$. The dimension of the null space is $p - m_{\neg\mathcal{A}}$ and thus one can solve an unrestricted linear regression for a vector of length $p - m_{\neg\mathcal{A}}$ after rotating the design matrix. This has a downside, however, of possibly making a sparse $\boldsymbol{X}$ into a highly dense one and thus increasing computational burden.

   The other strategy relies on characterizing the constraint in terms of fewer linearly independent restrictions. This can be done by decomposing $\boldsymbol{D}_{\neg\mathcal{A}}$ using, e.g., SVD or QR decompositions, left-multiplying by an orthogonal matrix, and then noting that there are $m_{\neg\mathcal{A}}$ binding restrictions. Then, the RLS approach by Greene and Seaks (1991) can be applied directly.

2. Clipping: This option is more crude. It says simply to set $E[1/\tau_k^2]$ to some large number, e.g. $1/\epsilon$ if $|\boldsymbol{d}_k^T\boldsymbol{\beta}| < \epsilon$. While slightly computationally faster, it may be more sensitive to the exact value chosen for ϵ as too large will run into problems of numerical instability. However, as $\boldsymbol{D}$ grows to be huge, this approach will be much faster as the cost of calculating either the null space or the simplified restrictions grows.

C.4 Adaptive LASSO

For the models used in the main text, I rely on the adaptive LASSO (Zou 2006), see Gertheiss and Tutz (2010) for an application to structured sparsity. I also use the weights suggested in Gertheiss and Tutz 2010 to normalize the size of the variables. This results in only a slight modification of the above results as it corresponds to multiplying each row of $\boldsymbol{D}$ by some weight $\hat{w}_k^\gamma$ where $\hat{w}_k$ is defined as follows where $\hat{\boldsymbol{\beta}}_{MLE}$ is some non-sparse estimator of the MLE. The kernel of the prior can be written as follows:

$$\hat{w}_k^\gamma = \left(\frac{1}{|\boldsymbol{d}_k^T \hat{\boldsymbol{\beta}}_{MLE}|} \right) p(\boldsymbol{\beta}|\lambda, \{\hat{w}_k\}, \gamma) \propto \left(-\lambda \sum_{k=1}^{K} \hat{w}_k^\gamma |\boldsymbol{d}_k^T \boldsymbol{\beta}| \right) \tag{44a}$$

The corresponding Gibbs Sampler is shown below, including for λ^2 if that is included in a Bayesian analysis.

$$1/\tau_j^2|\lambda, \gamma \sim \text{InverseGaussian}\left(\frac{\lambda \hat{w}_k^\gamma}{|\boldsymbol{d}_k^T \boldsymbol{\beta}|}, \lambda^2 [\hat{w}_k^\gamma]^2 \right) \tag{45a}$$

$$\lambda^2|\{\tau_k^2\} \sim \text{Gamma}\left(a_0 + [K+m]/2, b_0 + 1/2 \sum \tau_j^2 [w_j^\gamma]^2 \right) \tag{45b}$$

D Strategies for Estimating Regularization Strength

This section fleshes out some theoretical and practical details of selecting the regularization strength (λ) using the approaches outlined in the main text.

D.1 Cross-Validation

Conceptually, cross-validation is the most straightforward; see Hastie, Tibshirani, and Friedman (2009) for a general overview. For some data structures, however, this could be challenging; if one is attempting to control for geographic heterogeneity, a sensible strategy for cross-validation is to use stratified sample splitting based on those units so out-of-sample predictions can be obtained. This may lead to very limited information in the training set, however, and thus create sensitivity to the specific split chosen. However, if the data are such that cross-validation is feasible, it can be used with structured sparsity with no difficulty.

D.2 Information Criteria

A different approach is the large literature that relies on various information criteria (e.g. AIC or BIC) to select a model. The core idea is straightforward; one cannot simply evaluate the log-likelihood as a way of selecting models based on λ as smaller λ will invariably have better fits—as they are less sparse—at the cost of having a more complex model. One thus needs to have a statistic that evaluates the model's performance by penalizing the fit by some measure of the complexity of the model. In the case of general linear models, the commonly used AIC and BIC penalize based on the number of parameters in the model, adjusted by some scaling factor. Much research has been devoted to understanding the properties of these and related estimators in both the sparse and non-sparse case. Hui, Warton, and Foster (2015) provides a clear discussion of the developments in this literature, with focus on model selection with sparse underlying models.

Getting a measure of the model's complexity (or its "degrees of freedom") in the case of sparse models is more challenging. Two seminal papers (Tibshirani et al. 2004; Zou, Hastie, and Tibshirani 2007) proved that, for the linear model, an unbiased estimate of the degrees of freedom using the standard LASSO is the number of non-zero coefficients. From this result, the standard practice in the literature is to use this measure of the degrees of freedom for *non-linear* models (e.g. Park and Hastie 2007). The justification usually appeals to approximate normality of the objective function around the optimum. Some researchers have proposed biased-corrected measures of the degrees of freedom for the standard LASSO for non-linear models (Ninomiya and Kawano 2016) that could be used instead.

The crucial development for this paper, however, is the result in Tibshirani and Taylor (2012) where they derived an unbiased estimator of the degree of freedom, again assuming a linear model, for the *generalized* LASSO discussed in the first section of this paper. While the exact formula is outlined in their paper, it has a number of simple interpretations for common types of sparsity.[42] For example, if one is using temporal sparsity, the estimate of the degrees of freedom is the number of groups of fused coefficients. If one is using functional sparsity, it is the number of "knots" in the function. For categorical sparsity (i.e. geographic sparsity on a fully connected network), it agrees with the measure conjectured in Gertheiss and Tutz (2010) as the number of distinct fused

[42] If the design matrix ($\boldsymbol{X}$) is full rank, then the estimate is the nullity of the matrix formed by the rows of $\boldsymbol{D}$ where the restriction *does not hold*, i.e. $\boldsymbol{d}_k^T \boldsymbol{\beta} \neq 0$.

groups. Future work is needed to create a corrected estimator for non-linear models; however, again following existing practice, it seems reasonable to use their estimator of the degrees of freedom for non-linear models.

With that in hand, it is thus possible to vary λ over a grid and examine the evolution of one's preferred information criterion. This does require fitting over a range of values, like cross-validation, but only requires one fit per model and gives the coefficients of interest (i.e. estimated on the entire dataset) at the end. When doing this strategy, however, I suggest noting the λ suggested by multiple information criteria to show the robustness of one's results to the choice of λ.

D.3 Bayesian Prior

The above discussion is focused on the non-Bayesian paradigm. In the Bayesian case, there are two options. First, one can fix λ and use the Bayesian methods for inference on $\boldsymbol{\beta}$. Second, one can give λ a prior and use a fully Bayesian analysis. The later requires an exact characterization of the sparse prior, hence much of the discussion above. As noted before, it also is important to ensure that this prior is not "dominated" by the data and thus allow it grow in some way (Ratkovic and Tingley 2017, 2018). Mechanically, a Gamma prior on λ^2 has the nice property of being conditionally conjugate when using structured sparsity as Equation 46 shows.

$$p(\lambda^2|-) \propto (\lambda^2)^{(k+rank(\boldsymbol{D}))/2} \exp(-1/2 \sum_j \tau_k^2 \lambda^2) p(\lambda^2) \tag{46}$$

Thus, if λ^2 is given a Gamma prior, the posterior is Gamma. Inference here is straightforward, subject to the issue of calibrating the prior on λ^2.

E Details on Simulations

This section outlines more detailed results on the simulations in Section 5. First, it enumerates the methods used:

- SSp (AIC) - Structured Sparsity (AIC). Estimated using an adaptive LASSO penalty ($\gamma = 1$) with ridge-stabilized weights. λ was chosen by picking the model with the best AIC over a grid of 20 λ ranging from 0.01 to 10 at equal increments on a logarithmic scale. I use a fully

connected structure.

- SSp (BIC) - Structured Sparsity (BIC). Estimated using an adaptive LASSO penalty ($\gamma = 1$) with ridge-stabilized weights. λ was chosen by picking the model with the best BIC over a grid of 20 λ ranging from 0.01 to 10 at equal increments on a logarithmic scale. I use a fully connected structure.

- FE - Fixed Effects. Estimated using `glm` and an interaction of indicator variables for each group with the treatment, i.e. `glm(y ~ x + d * g)`.

- RE - Random Effects. Estimated using `glmer` (Bates et al. 2015) and a random slope for the effect of treatment, i.e. `glmer(y ~ x + d + (d | g))`

- BayesGLM - From `arm`; a generalized linear model with a prior on each coefficient from Gelman et al. 2008 to avoid separation.

- LASSO - λ chosen using ten-fold cross-validation from `glmnet` (Friedman, Hastie, and Tibshirani 2010); the formula is `cv.glmnet(y ~ x + d * g)`

- ENet1 - Elastic Net with $\alpha = 0.50$ following Grimmer, Messing, and Westwood (2017). λ chosen using ten-fold cross-validation from `glmnet`. Same formula as LASSO.

- ENet2 - Elastic Net with $\alpha = 0.25$ following Grimmer, Messing, and Westwood (2017). λ chosen using ten-fold cross-validation from `glmnet`. Same formula as LASSO.

- FindIt - Estimated using default settings in Imai and Ratkovic (2013).

- SVM - Estimated using polynomial kernel following Grimmer, Messing, and Westwood (2017) and package `e1071` (Meyer 2019). Default settings were used.

- BART - Estimated using default settings from `BART` package (Sparapani, Spanbauer, and McCulloch).

- RF - Estimated using a forest of 1,000 trees with $\lfloor G/3 \rfloor + 2$ variables drawn per tree. Estimated using the `randomForest` package (Liaw and Wiener 2002).

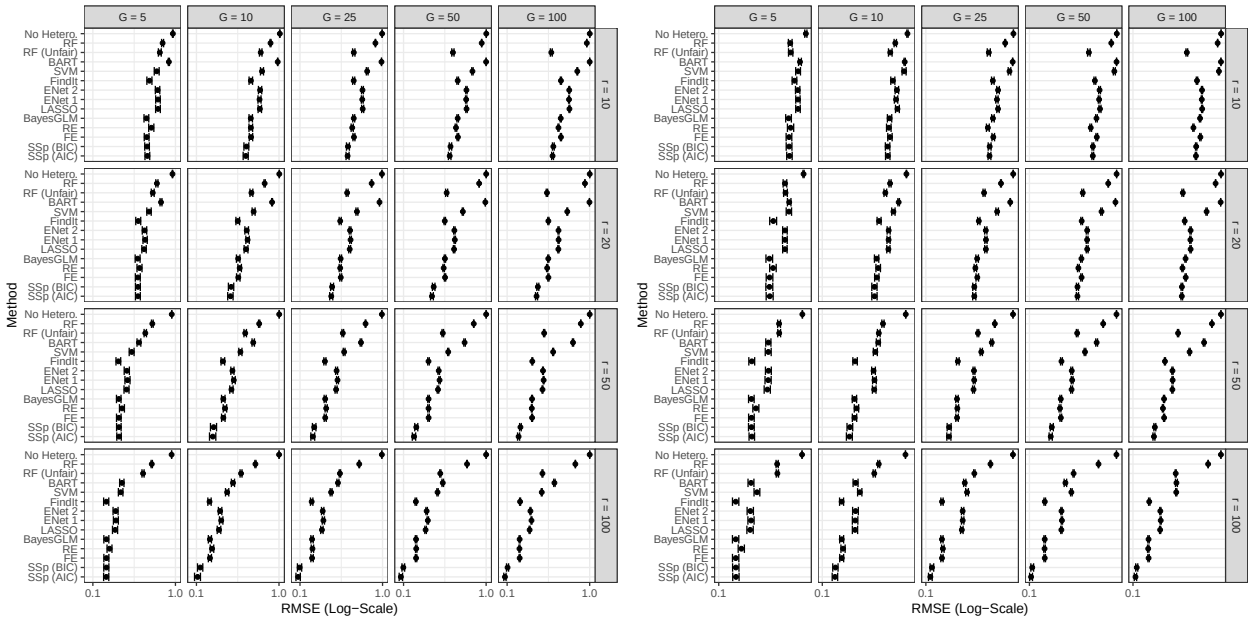
- RF (Unfair) - Estimated using a forest of 1,000 trees with all three variables (confounder x, treatment, and a categorical variable for G) included. This creates splits based on the unordered categorical G based on, at each split, ranking the categories in terms of the observed outcome and using that variable to decide which groups to allocate. It is denoted as "unfair" for reasons discussed in the main text; specifically, that it uses the outcome to create variables to use when predicting the outcome! Estimated using the `randomForest` package (Liaw and Wiener 2002).

- No Hetero. - No Heterogeneity. A model estimated with no heterogeneous effects, i.e. `glm(y ~ x + d)`.

Second, it shows the full results for all methods across all simulation environments (i.e. varying $\mathcal{S}$ and the outcome type). The major point to note is that, broadly, the results hold up as G varies; for $G > 5$, structured sparsity (selected either via the AIC or the BIC) is usually the best performing model for the linear case where the truth is mostly grouped. For the mostly sparse case, the same pattern holds (structured sparsity is usually the top-performing model) although sometimes the "unfair" random forest (RF - Unfair) is the best performing method.

Figure 4: Simulations for All Methods: Linear Data Generating Process

(a) Mostly Grouped Effects ($\mathcal{S} = \lfloor G/2 \rfloor$) (b) Mostly Sparse Effects ($\mathcal{S} = \lfloor G/4 \rfloor$)



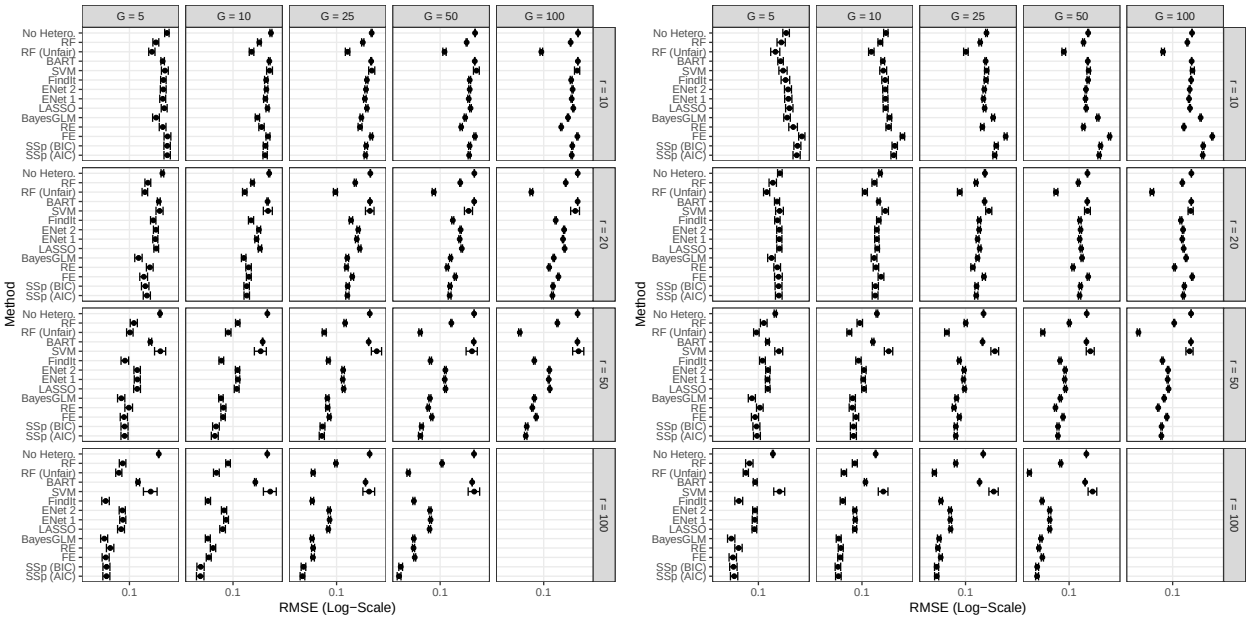
Finally, for the models with a binary outcome, the results are shown below. The generative model in that case uses the same linear predictor $(x_i + \tau_{g[i]} d_i)$ but generates the outcome as follows:

$$y_i \sim \text{Bernoulli}\left(\frac{\exp(x_i + \tau_{g[i]} d_i)}{1 + \exp(x_i + \tau_{g[i]} d_i)}\right)$$

The results for those simulations are shown below. For the mostly grouped case ($\mathcal{S} = \lfloor G/2 \rfloor$), structured sparsity usually performs the best although is sometimes out-performed by random effects (especially for small G or r) or the unfair random forest. It is roughly in the middle of the pack of the methods for most values in the mostly sparse case as noted in the main text.

Figure 5: Simulations for All Methods: Logistic Data Generating Process

(a) Mostly Grouped Effects ($\mathcal{S} = \lfloor G/2 \rfloor$) (b) Mostly Sparse Effects ($\mathcal{S} = \lfloor G/4 \rfloor$)



F Details on Giger and Klüver 2016

This section provides details on the re-analysis of the models in Giger and Klüver (2016).

F.1 Discussion of Control Variables

First, to create the measure of cross-pressure, I use the logical implications in Table 3. This shows that if the variable "defect" and "party congruence" have the same value, the MP must be in a

cross-pressured situation.[43]

The other control variables are included exactly as in Giger and Klüver (2016). They are listed below; see the original paper for details. All the main specifications included random effects for party and canton.

> months until next election, number of MPs per canton, closeness of referendum, closeness of parliamentary decision, salience, referendum type (obligatory, facultative, initiative)

F.2 Estimating Structured Sparsity: Cross-Validation

I estimate a model using structured sparsity that is a slight modification of the original model in Giger and Klüver 2016. I will rely on structured sparsity that uses the adaptive LASSO weights to ensure the oracle property and thus I need to use a consistent estimator of the true coefficients. However, given that the amount of data is small for some cells, and thus the point estimates may be rather noisy, I rely on a ridge-stabilized version of this model ($N(0, 2)$ prior on each interactive coefficient). I set $\gamma = 1$ and use the finite-sample weights suggested in Gertheiss and Tutz (2010).

To decide the optimal λ, I fit each of the three structures over a grid of λ where the model is run for a short period of time (1,000 iterations with 1,000 burn-in) to limit the computational burden of (coarsely) choosing λ. I then calculated the WAIC at each λ. The plot for each structure and each data type (cross-pressured or consensus) is shown in Figure 6.

As expected, each shows the u-shaped relationship expected with cross-validation where a moderate amount of sparsity outperforms either a model with no sparsity or with maximal sparsity. The Referendum structure does noticeably worse than the Lattice or Fully Connected structures, and the latter two perform similarly.[44] To ensure the estimated groups are more coherent, I rely on the model that uses the Lattice structure for the main results.

To estimate the models in the main text, I then run a much longer sampler to fully explore the posterior: three chains each with 5,000 iterations and 5,000 burnin with over-dispersed starting

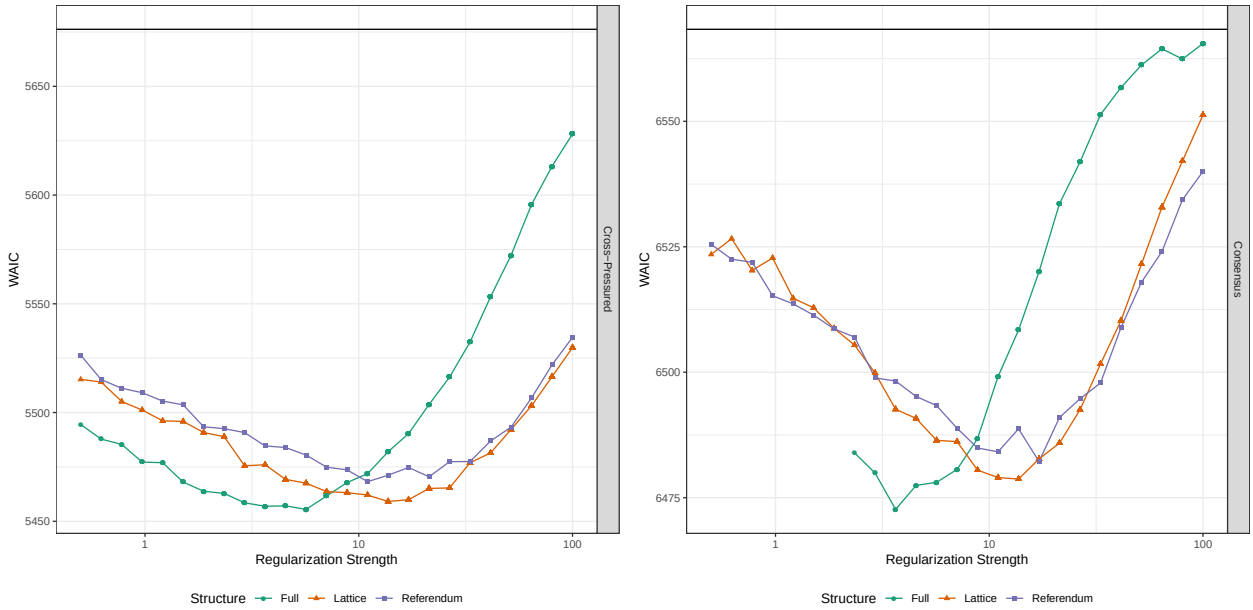
[43]For example, if the MP defects and agrees with their party, then the party must disagree with the local. If the MP defects and, in doing so, does not agree with their party, then there must have been a consensus position that MP rejects. Etc.

[44]For robustness, I ran each of these structures using the longer sampler described below. In those specifications, the difference between the WAIC for each of those two structures is less than one for cross-pressured and consensus votes.

Figure 6: Bayesian Cross-Validation

(a) Cross-Pressured Votes

(b) Consensus Votes



Note: Each figure shows the WAIC from a Bayesian cross-validation where the model is fit for each regularization strength (λ). Lower values indicate better performance; the horizontal line indicates the fit of the original model (i.e. with no heterogeneity).

values.[45] The model converges rapidly as suggested by the Gelman-Rubin diagnostic (none above 1.03) and the Geweke diagnostic (95% between 0.268 and 1.43), with the largest being 1.91. This supports other results that the Bayesian LASSO and Polya-Gamma samplers show good mixing (e.g. Kyung et al. 2010; Polson, Scott, and Windle 2013). Pleasingly, the WAIC estimated from this long run is very close to the WAIC estimated from the short run used to select λ.

F.3 Details of Groups

I list here the members of the non-singleton groups (i.e. groups with more than one member) that are also not part of the main group (i.e. 30+ members) described in Figure 2.

F.4 Comparison of Uncertainty Across Methods

Finally, I show the estimated posterior intervals for random effects and structured sparsity side-by-side. It confirms the following stylized fact: Both methods have broadly similar results in terms

[45] I take the weights for the adaptive LASSO, i.e. the (ridge-stabilized) maximum likelihood solution and multiply each coefficient by a number drawn at random from $Unif(-2, 2)$.

Table 7: Disaggregated Groups from Structured Sparsity

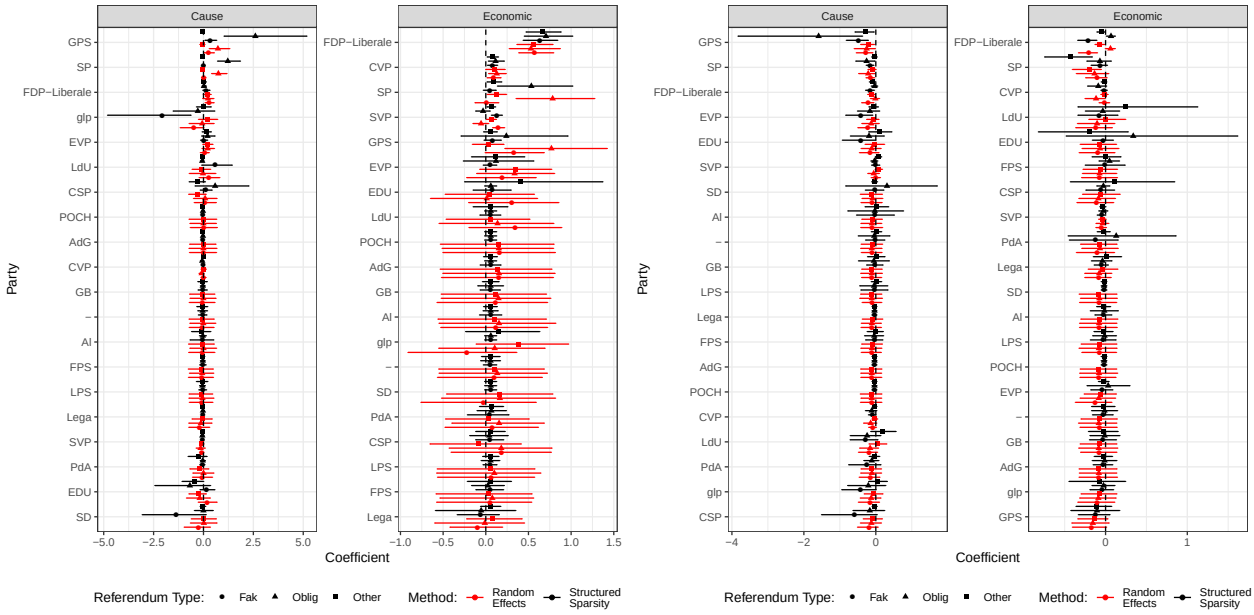
Vote Type	Group Name	Interest Group	Members
Consensus	Misc. Facul (1)	Cause	glp Facul, CSP Facul, EDU Facul, EVP Facul, GPS Facul
Consensus	Misc. Other	Cause	Al Init, glp Init, EDU Init, SVP Init
Consensus	Misc. Facul (2)	Cause	PdA Facul, LdU Facul, FDP-Liberale Facul, SP Facul, SP Oblig
Consensus	Misc. Oblig	Cause	CVP Facul, glp Oblig, CSP Oblig, EDU Oblig, PdA Oblig, LdU Oblig, EVP Oblig, CVP Oblig
Consensus	Misc. Oblig	Economic	EDU Oblig, PdA Oblig, FPS Oblig, EVP Oblig, FDP-Liberale Oblig
Cross-Pressured	Misc. Facul	Cause	EDU Facul, FDP-Liberale Facul, CSP Facul
Cross-Pressured	Misc. Christian	Cause	EVP Init, EVP Oblig, CSP Oblig
Cross-Pressured	Misc. Init	Cause	PdA Init, CSP Init
Cross-Pressured	EDU	Cause	EDU Init, EDU Oblig
Cross-Pressured	FDP	Cause	FDP-Liberale Init, FDP-Liberale Oblig
Cross-Pressured	Lega	Economic	Lega Facul, Lega Oblig
Cross-Pressured	FDP	Economic	FDP-Liberale Facul, FDP-Liberale Init, FDP-Liberale Oblig

of which sub-units effects are statistically distinguishable from zero. However, they differ quite dramatically in the effect given to groups with limited or no variation. The random effect model performs in the expected way; in the absence of variation, it gives an effect that is centered the global effect plus some lafallrge random error. By contrast, structured sparsity gives quite precise—small–values to those observations. This is because it leverages the fact that they should be shrunk towards their neighbors—for whom data *is* observed—and thus generates smaller uncertainty around the estimated effect.

Figure 7: Comparing Heterogeneous Effects by Method

(a) Cross-Pressured Votes

(b) Consensus Votes



Note: The 95% credible interval on each coefficient is shown with the posterior mean indicated by a marker.Each panel shows the effect for cause and economic interest groups separately. Red lines indicate structured sparsity; black indicate random effects.

G Details on Timing

This section conducts some simple simulations of "like-for-like" tests of the EM algorithm derived above and a major existing approach (ADMM; Zhu 2017) for finding the solution to a structured sparse prior with only $\boldsymbol{D}$. To facilitate a like-for-like comparison, I look at cases that find estimates for a single λ. See Zhu (2017) for comparisons of their method against the path algorithms in Arnold and Tibshirani (2016).

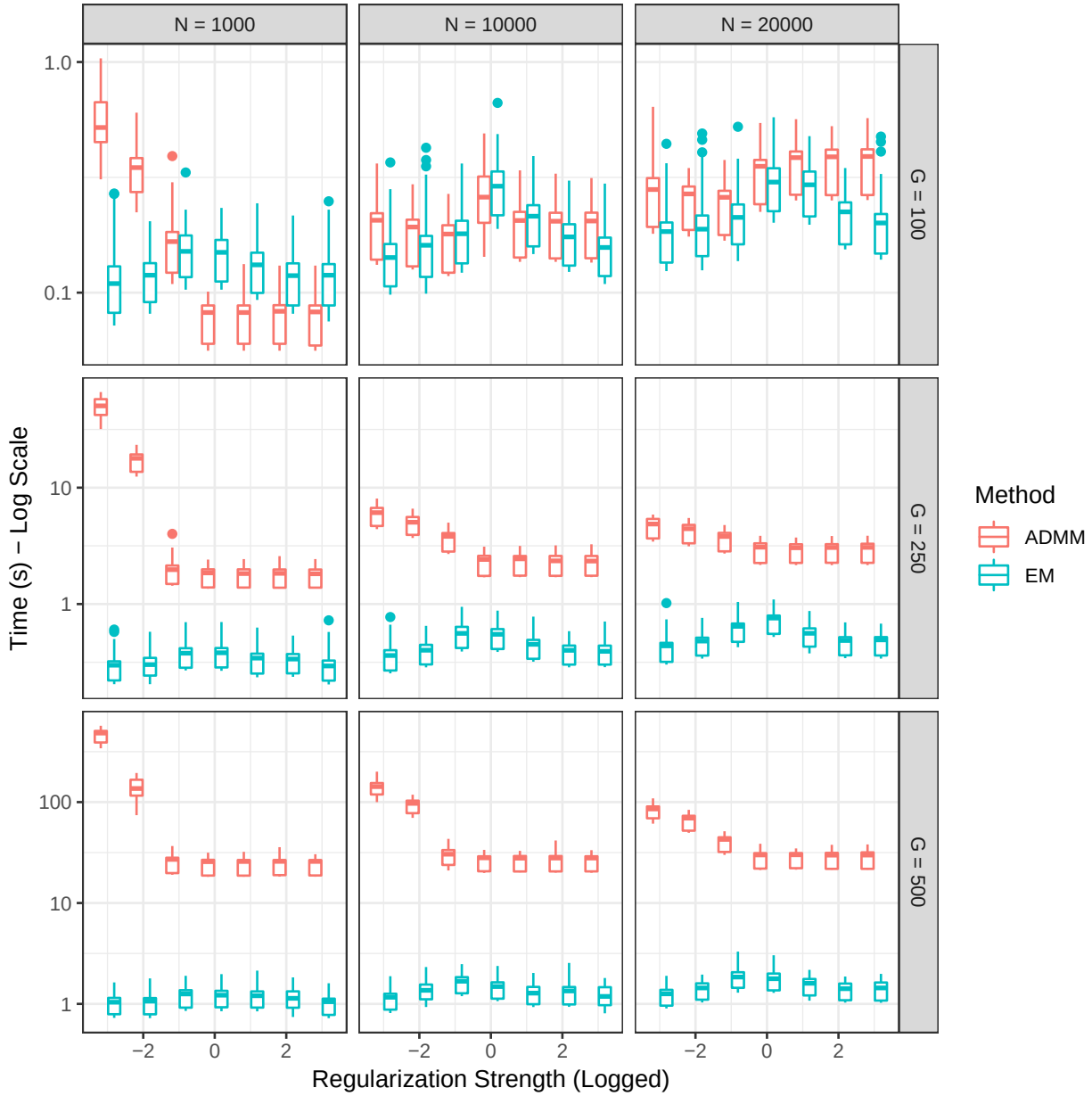
I used a simple environment that mirrors the simulations in the main text:

- Data: Allocate N observations i to G groups (indexed by j) at random.
- Coefficients: Simulate $\beta_j \sim N(0, 1)$.
- Outcome: $Y_i \sim N(\beta_{j[i]}, 1)$. Standardize to have mean-zero and variance-one, as ADMM cannot estimate the error variance of the linear regression and thus this is set fixed to $\sigma^2 = 1$ in both models. This ensures that the regularization strength is constant across models.

In the simulations, I vary three key parameters that affect the speed of inference. First, I vary $N \in \{1000, 10^4, 2 \cdot 10^4\}$. I then vary the number of groups (coefficients to estimate; shaping the dimensionality of $\boldsymbol{D}$) in $G \in \{100, 250, 500\}$. Finally, I vary the sparsity of the model, i.e. $\lambda = 10^{-a}; \quad a \in \{-3, -2, \cdots, 2, 3\}$ as the models that are more sparse can be estimated more quickly (insofar as both procedures quickly realize that most $\beta_j \approx 0$ and thus convergence is rapidly achieved). I estimate models with categorical sparsity, i.e. a fully connected network between all β_j and thus $\boldsymbol{D}$ has dimensionality $\binom{N}{2}$ by G. I use the default settings in ADMM for stopping values (absolute tolerance of 10^{-4} and relative tolerance of 10^{-2}); I set the EM algorithm to comparable standards (the largest change in any element of $\boldsymbol{\beta}$ is below 10^{-4}; a relative change in the log-likelihood below 10^{-4}).

Figure 8 shows the results; they confirm the benefits of using an EM approach as the dataset grows in size. Please note that the vertical axes are logged but also different for each row (G). We see that ADMM has an advantage when G is small, although both methods converge in mere seconds. By contrast, as G grows, ADMM becomes noticeably slower especially for models that are somewhat dense (i.e. smaller regularization strength). Thus, for small cases, ADMM is somewhat faster, but for larger cases that may easily occur for applied data analysis, the EM procedure derived in this paper is markedly faster.

Figure 8: Speed of Estimation Procedures (in Seconds)



Note: This figure showns the distribution of the run-time in seconds by each method, when varying the key parameters discussed in the main text. It shows a standard boxplot; N is the number of observations; G is the number of groups (i.e. the number of variables to be estimated). The regularization strength is on the horizontal axis and is 10^{-a} where the number is that shown on the axis. Please not that the vertical axis uses a log scale to show comparability and that the axis values differ as G varies, i.e. each row of figures has a differently scaled vertical axis.