Modelling Heterogeneity Using Complex Sparsity

Max Goplerud*

December 16, 2018

Abstract

Data analysis starts by making theoretically informed assumptions about the structure of the world. Yet, we also typically have reasons for expecting certain types of heterogeneity in our analysis (e.g. temporally or spatially varying effects). This paper provides a unified framework for modelling the most common sorts of heterogeneity by leveraging the concept of 'complex sparsity' from machine learning. To implement this idea for social scientific questions, it is necessary to develop a new inferential procedure as existing methods cannot quantify uncertainty nor estimate non-linear models without approximations. This paper addresses this by deriving a new Bayesian representation of complex sparsity and associated inferential procedures. I apply this framework to a recent paper on segregation and public goods in the United States (Trounstine 2016), focusing on the usefulness of complex sparsity in estimating regression-based fixed effects models. I find that while the original specifications in the paper are not robust to usual methods for addressing unobserved confounding, using complex sparsity strengthens the paper's original results.

1 Introduction

Analyzing data in social sciences begins by making assumptions about the world. These assumptions have a variety of names including, but not limited to, 'no omitted variables', 'conditional ignorability, and 'linearity'. Researchers justify these assumptions by either appealing to some property of how the data are generated and by using our theoretical knowledge of the question at hand. With these assumptions, it is possible to estimate our quantities of interest using this simplified representation of the world as we rarely believe we are in the 'worst of all possible worlds' where all units have highly idiosyncratic relationships between independent and dependent variables.

However, researchers often impose functional form assumptions in our models—not necessarily because they believe them, but because they lack a principled and robust way of relaxing them that does not undermine their ability to draw reliable conclusions. For example, a vast amount of data occurs in an (implicitly) time-series context: A straightforward approach to a robust analysis

*Draft. `http://mgoplerud.com/papers/Goplerud_Sparsity.pdf`; I thank participants at the American Political Science Association 2018, PolMeth 2018, and Text-As-Data 2018 for helpful comments. I give special thanks to Gary King, Kosuke Imai, Walter Mebane, Brandon Stewart, Marc Ratkovic, Dustin Tingley, and Jeremy Wachter for helpful comments on earlier versions.

would interact all covariates with dummies for time, i.e. run period-by-period regressions, but this will likely undermine one's ability to draw inferences by exploding the number of covariates and leading to large standard errors (over-fitting or under-powering our results). Further, many datasets have a geographic dimension; it is often a heroic assumption to assume that the effects of variables are stable across space, but again relaxing this assumption by interacting variables with dummies for geography will lead to severe inferential issues. Given those concerns, the most common approach is to assume stability for most variables—pooling information across time or space.

This problem is also pressing for scholars estimating causal effects. After one has defended the identification strategy (randomization, fixed effects, etc.), it is necessary to *estimate* the causal quantity of interest. Typically, we target an average causal effect—averaging across time, space, and unit heterogeneity. This procedure allows one to look for treatment effect heterogeneity by using our theoretical knowledge without dramatically limiting the power of our analysis. If, for example, one expects there to be heterogeneity over space, but we are willing to assume that neighbors are likely similar, then we can quite flexibly estimate spatially varying treatment effects using this framework. In a fixed effects context, the literature argues that we should include fixed effects at the 'smallest' unit for which there is variation in treatment; for example, one might include state-fixed effects instead of city-fixed effects. However, if we were willing to assume that the fixed effects likely do not change often neighboring cities, this framework would allow us to collapse the fixed effects into groups and thus improve power and efficiency.

Despite having very different provenances, this paper suggests a unified framework to these problems by leveraging the idea of 'sparsity' from statistics and machine learning. In its canonical form, 'sparsity' is the idea that when doing a regression with many covariates (possibly more than the number of observations), it is plausible to 'bet' that many of them are zero and can be safely ignored in the analysis (Hastie, Tibshirani, and Friedman 2009). We can develop estimation procedures that use the data to select which variables are non-zero, given our theoretical prior about a 'sparse' world. This paper argues that, if we conceptualize sparsity more broadly, it can exactly represent most of the functional form assumptions that we make in our analyses. Thus, rather than imposing homogeneity by fiat, we can use our theoretical knowledge to 'bet' on stability—but estimate a model that allows for heterogeneity when it is relevant.

In the case of geographic analysis above, this would be equivalent to assume that from city to neighboring city, most coefficients are stable (do not change) but sometimes there are changes.

The researcher would need to defend that assumption on theoretical grounds, but can remain agnostic on exactly which variables change, when those changes occur, etc. and thus estimate a more flexible model for addressing this heterogeneity. Perhaps surprisingly, this can be done without running into the problems of over-fitting and lack of power noted above by leveraging the fact that methods to estimate complex sparsity are *exactly* designed to deal with the case where the number of parameters is very large. As discussed in more detail later, many of these procedures for inducing a sparsity model have the asymptotic property of recovering the 'true' model with correct standard errors.

The key challenge in implementing these models is two-fold. First, as many of the inferential procedures have their origins in computer science, they are not suited for social scientific inquiry. Specifically, most are predicated on a linear model or particular data structures that do not arise for social scientific questions.[1] While approximations exist for non-linear models (e.g. Fan and Li 2001), this is mostly focused on prediction for the canonical sparse case. As we typically care about inference, it is important to have techniques that exactly optimize the likelihood function of interest. Further, again coming from the focus on prediction, there is no easy and straightforward to accurately quantify the uncertainty in those models. Finally, by focusing on perhaps the most famous way of inducing sparsity (the LASSO - Tibshirani 1996), their inferential procedures are saddled with the well-known problems of the LASSO (bias in the non-sparse coefficients) and are not easily transferable to the more sophisticated types of regularization designed to address these problems. While fine for purely predictive questions, this is not appropriate when we are attempting to do inference based on the coefficients themselves.

Fortunately, these canonical sparse models have been translated into Bayesian frameworks that address most of these problems (Park and Casella 2008; Kyung et al. 2010). However, the Bayesian turn has one major disadvantage: It is much more computationally demanding to estimate these models—enough to make them intractable for the normal applied researcher. Indeed, most of the modern Bayesian methods for inducing sparsity ('global-local shrinkage'; Polson and Scott 2011a) are rather computationally challenging to estimate versus the simple LASSO.[2]

This paper solves these problems by leveraging the Bayesian representation of sparsity to derive a fast, scalable, and exact method of finding the (penalized) maximum likelihood estimate or the posterior mode using an EM algorithm (Dempster, Laird, and Rubin 1977). One can use that

[1] Specifically, when the design matrix is the identity matrix (Tibshirani et al. 2005), but see Tibshirani and Taylor (2011) for a method that works for a full rank design matrix.

[2] For example, the horseshoe prior is notoriously difficult to work, e.g. trouble with mixing as noted by Johndrow, Orenstein, and Bhattacharya (2018) or reliance on more complicated samplers as in Carvalho, Polson, and Scott (2010), with despite having good theoretical properties.

point estimate for prediction or, if measures of uncertainty are desired, it can be used to start the fully Bayesian sampler—itself simple to estimate—and thus achieve rapid convergence. This framework will work for a wide class of types of sparsity (including all those outlined above) as well as allowing the researcher to arbitrarily 'mix and match' types of sparsity depending on their particular data structure. By employing other work in Bayesian data augmentation, this framework also applies to the most common dependent variables found in our research (linear, binary, count, ordinal, and multinomial) without requiring approximations. To address the issues with the LASSO penalty, I show that the fast framework for inference can be adapted to include other forms of sparsity including the adaptive LASSO, LASSOplus (Ratkovic and Tingley 2017), and other existing global-local priors.

The insight behind this framework is the following: Conditional on data augmentation, explained in detail later in the paper, these complex types of sparsity can be cast as a mixture of normals and thus iteratively weighted least squares can be used to find a posterior mode. In many cases, if we condition on the strength of regularization and have a 'normal' likelihood function, this is the unique posterior mode. For a very large class of complex sparsity, estimation can be made even faster by noting that they can be written as dynamic or sparse linear models for which a large amount of pre-existing theory allows us to solve rapidly. Thus, by iterating between augmentation and maximization, I show that an EM algorithm (Dempster, Laird, and Rubin 1977) can be used to deterministically find a posterior mode. Overall, with these results in hand and with the accompanying package, it is possible for social scientists to estimate more plausible models by incorporating their explicit assumptions about the complex types of sparsity in their data—and thus allowing heterogeneity from that assumption to appear.

I apply this framework to estimating models with fixed effects, focusing on the most common approach of estimating these via simple regression adjustment.[3] The main trade-off in this approach is that while this is a causally credible strategy, including fixed effects at the smallest level of aggregation for which there is variation on the treatment variable is a very data-intensive strategy. Often, the data does not have much variation at that level and thus the associated estimates are very noisy, limiting the ability to draw reliable conclusions. I suggest that complex sparsity provides a way of increasing the efficiency of fixed effects point estimates if the researcher is willing to assume (and defend) that there is likely some structure in the underlying heterogeneity.

Specifically, I focus on a recent paper by Trounstine 2016 that examines the role of residential

[3]See Imai and Kim (Forthcoming) for a discussion of the downsides of this approach and different inferential strategies based on matching. The results here can also be applied to that framework insofar as it can be cast as a (weighted) linear model with unit-specific dummy variables, e.g. estimation of unit-varying time trends.

segregation in American cities on political outcomes and public goods provision. I show that complex sparsity can help bolster the results in two ways; first, re-examining the results on political outcomes shows that the reported specification does not effectively control for unit-level unobserved confounding. Doing so correctly via classic fixed effects approach leads to a point estimate that is underpowered and insignificant. Complex sparsity, however, shows that even weak assumptions of groups in the underlying heterogeneity leads to clearly statistically significant effects. Second, I examine the robustness of the results on public goods provisions to time-varying unobserved heterogeneity. I show that, again, the typical approach of including unit-specific time trends leads to a smaller and insignificant result. Complex sparsity, assuming that many units shared the same underlying unobserved time trend, agrees that the effect is noticeably smaller than the one reported in the original results, but it retains statistically significance.

2 The Flavors of Sparsity

Any overview of sparsity should probably start by discussing the LASSO (Tibshirani 1996). Fully known as the 'Least Absolute Shrinkage and Selection Operator', it can be most straightforwardly thought of as a way to estimate a penalized maximum likelihood model. Instead of simply finding the set of coefficients $\boldsymbol{\beta}$ that maximize the likelihood, one finds the $\boldsymbol{\beta}$ that maximize the likelihood *minus* a penalty term that penalizes more 'complex' models. For the LASSO, the penalty term is $\sum_j \lambda|\beta_j|$ for each coefficient in $\boldsymbol{\beta}$. The power of this penalty is that, at the optimum, many of the coefficients would be set to exactly zero and thus induces sparsity in its results. Further, the LASSO can be estimated when there are more variables than observations. From this innovation, a huge cottage industry of flavors of LASSO has grown in the computer science and statistics literature. Table 1 synthesizes some of the major variants, especially those from the original creators of the LASSO and their co-authors, that are especially relevant to social science questions.[4] The original ('vanilla') LASSO penalizes single coefficients to be zero. This enforces the common idea of 'sparsity'. The next four restrictions are all particular instantiations of the 'generalized LASSO' (Tibshirani and Taylor 2011) that penalize linear combinations of coefficients; different linear combinations have different properties, and thus I outline some common sub-types of complex

[4]This is not exhaustive; some other important variants are the group LASSO (Yuan and Lin 2006), the graphical LASSO (Friedman, Hastie, and Tibshirani 2008), the elastic net (Zou and Hastie 2005), and the adaptive LASSO (Zou 2006). The (naive) elastic net is trivially included in this framework by recasting the quadratic (ridge) penalty as part of the transformed covariate matrix. The group LASSO is discussed extensively in other work by the author (Goplerud 2018b) and briefly integrated into this framework in Appendix B. Incorporating the graphical LASSO into this framework is an interesting area for future research. The adaptive LASSO, and its variants, are trivially included by modifying λ by variable or group specific weights as outlined in later sections.

sparsity that are important enough to warrant their own names. 'Temporal' sparsity is the idea that coefficients should be fused to some immediately prior value (the 'fused' LASSO; Tibshirani et al. 2005). 'Functional' sparsity is the idea that coefficients should be fused in some polynomial way; for example, 'first-order functional sparsity' would create a piece-wise linear relationship between the coefficients. As Table 1 shows, we see that *if and only if* the three $\beta_{j,t}$ lie on a line, is the penalty zero. Higher order functional sparsity is designed to mimic the behavior of smoothing splines without having to select complicated tuning parameters such as knots in advance (Kim et al. 2009; Tibshirani and Taylor 2011; Tibshirani 2014).[5] 'Geographic' sparsity is the idea that neighboring geographic units should have their coefficients fused together; this would suggest that the effect of a covariate may vary county-by-county, but that neighboring counties should have similar values. 'Categorical' sparsity is the idea that we want to shrink all elements of $\boldsymbol{\beta}_j$ to all other elements, i.e. a network where every element of $\boldsymbol{\beta}_j$ is connected.[6] This would be the way to take a many-leveled categorical variable and attempt to find clusters of levels with the same effect by shrinking them all together.

Beyond the generalized LASSO, the major other conceptual innovation is to force *groups* of coefficients to be all zero, e.g. contrast-coded categorical variables (Yuan and Lin 2006), and more generally quadratic combinations of variables to be zero; this is known as the 'group LASSO'. This has the benefit of shrinking entire groups of coefficients to zero; this would be useful if one wanted to say, exclude a categorical variable entirely.[7] Detailed exploration of this is left to future research, but Appendix B outlines how it can easily be integrated into this framework.

From this table, we already see that the combination of the generalized LASSO and its particular instantiations and the group LASSO give us many types of theoretically relevant sparsity that we would want to model in our data. However, the framework outlined in this paper allows for us to go further: Some particularly interesting types of sparsity can be created by mixing

[5]The exact formula for a penalty of some polynomial order k is as follows (Tibshirani and Taylor 2011). For some $\beta_{j,t}$ where $t \leq T-k-1$,

$$\lambda \left| \sum_{s}^{t+k+1} (-1)^{s-t} \binom{k+1}{s-t} \beta_{j,s} \right|$$

For example, second-order polynomial sparsity would have $\beta_{j,1} - 3\beta_{j,2} + 3\beta_{j,3} - \beta_{j,4}$ inside the absolute value function.

One can think of '0-order' functional sparsity as equivalent to the temporal LASSO. Further, '-1-order' functional sparsity gives the same penalty as the vanilla LASSO.

[6]In effect, one can think of both geographic and categorical sparsity as such: There is an undirected graph where the nodes are the set of coefficients. Any connected nodes should be penalized to be equal, i.e. the absolute value of the difference of the coefficients. Geographic sparsity connects only those adjacent nodes; categorical sparsity is a fully connected network where all coefficients are connected to each other.

[7]As Simon et al. (2013) note, however, if some group was not pulled to zero, there would be no sparsity penalty on the individual coefficients; thus, they propose a combination penalty of the group LASSO and original LASSO sparsity (vanilla). The framework in this paper admits an arbitrary mixture of generalized and group sparsity.

Table 1: Taxonomy of Sparsity

Name	Penalty	Effect	Examples
Vanilla	$\lambda\|\beta_j\|$	Shrink to Zero	Classification with Text Data
Temporal	$\lambda\|\beta_{j,t} - \beta_{j,t-1}\|$	Shrink to Past Value	Change Point Detection
Functional	$\lambda\|\beta_{j,t} - 2\beta_{j,t+1} + \beta_{j,t+2}\|$ $(k = 1)$	Shrink to Polynomial	Non-Linearities in Explanatory Variables
Geographic	$\lambda\|\|\boldsymbol{G}\boldsymbol{\beta}_j\|\|_1$	Shrink to Neighbors	Geographic Heterogeneity
Categorical	$\lambda\|\|\boldsymbol{C}\boldsymbol{\beta}_j\|\|_1$	Shrink to All Others	Collapsing Categorical Variables

Selected 'Mixed' Variants

(1) Temporal Sparse LASSO $\quad \lambda_1|\beta_{j,t}| + \lambda_2|\beta_{j,t} - \beta_{j,t-1}|$

Note: $\boldsymbol{G}$, $\boldsymbol{C}$. denote matrices to induce the stated pattern of sparsity. $\boldsymbol{G}$ and $\boldsymbol{C}$ are particular examples of the generalized LASSO (Tibshirani and Taylor 2011) where $\boldsymbol{G}$ includes one row for every pair of neighboring units and $\boldsymbol{C}$ includes one row for every pair of elements of $\boldsymbol{\beta}_j$. $\boldsymbol{\alpha}$ represents an observation-specific parameter added to robustify against outliers. See the text for a discussion.

the types together (e.g. 'temporal sparse LASSO', i.e. shrink many coefficients to zero and to their past values) as other cases of the generalized LASSO.[8] As there are nearly infinitely many relevant types of sparsity to impose, I focus on the generalized case in the main text for clarity of exposition; Appendix B outlines the specifics of the other cases and discusses some details of implementation.

The key problem, however, when examining the richness of the flavors of sparsity is that inference—even in the simple LASSO case—is rather challenging. After iterating through a variety of techniques, a very popular estimation framework for estimating the vanilla LASSO is cyclical coordinate descent (Friedman et al. 2007) and is implemented in a fast `R` package `glmnet` (Friedman, Hastie, and Tibshirani 2010). This method that optimizes one element of $\boldsymbol{\beta}$ at a time—holding all other elements constant—and cycles through the elements of $\boldsymbol{\beta}$ repeatedly until the objective function has converged.[9] Using results from Tseng (2001), it can be shown this method is guaranteed to converge as the problem can be written as a (concave) likelihood function $\ell(\boldsymbol{\beta})$ plus a series of separable convex penalties $h_j(\beta_j)$—even if they are not differentiable:

$$\ell(\boldsymbol{\beta}) - \sum_j h_j(\beta_j) \tag{1}$$

[8] Other types of sparsity, however, can be developed by combining the group LASSO with the more traditional types of sparsity. Consider the group-sparse temporal LASSO: Imagine that we were attempting to predict vote intention using the seven-point partisanship scale. A group-sparse temporal LASSO penalty would encourage (i) all six dummy variables to be equal to their prior values; (ii) even if all six are not fused, the individual dummy variables are encouraged to be equal to their prior values. Alternatively, one could imagine a 'group functional LASSO' that encouraged the trends in each of these coefficients over time to be linear. See Goplerud (2018b) for discussion of these cases.

[9] Another popular method in computer science, OWL-QN, relies on a quasi-Newton procedure (Andrew and Gao 2007).

Note, however, that all of the more complex cases violate this rule as multiple β_j are 'linked' together in the penalty functions. Thus, for a cyclical algorithm to work in a straightforward fashion, it should optimize all of the linked coefficients—i.e. all of the $\boldsymbol{\beta}_j$ in a group such as all of the $\beta_{j,t}$ with temporal sparsity. Modified versions of the cyclical algorithm have been proposed for particular cases of the linear model; for example, Friedman et al. (2007) discuss estimation algorithms for the temporal LASSO. However, they note that their framework is not guaranteed convergence for a general regression case and thus is likely not fit for purpose for general social science research. Tibshirani and Taylor (2011) provides an algorithm for the generalized LASSO, including the temporal LASSO as a special case, but notes that their procedure as well runs into trouble in the case when we have more variables than observations (or the design matrix is not full rank) and requires a rather *ad hoc* fix.[10] Further, as the combinations of penalties grow more complex and possibly differ across variables, the existing cyclical approaches or other forms of convex optimization require careful tailoring to the particular case at hand. Moreover, optimization algorithms may exist, they are not designed to have the rapid speed of the cyclical method and thus scale much less well as the size of the data grows.[11] Finally, their frameworks are only suited for the generalized LASSO; this may, however, be fruitfully combined with the group LASSO and, in this case, their algorithms are not suitable for inference. Oelker and Tutz (2017) discusses an alternative framework for estimating the generalized and group LASSO, although inference is again based on approximations (even in the linear model).

3 Inference Using The Bayesian Turn

I tackle this problem by this problem by leveraging the Bayesian formulations of the (vanilla) LASSO. Instead of thinking of the LASSO as a penalty on a likelihood function to ensure a sparse solution, we can think of it as imposing a prior on the coefficients that corresponds to a posterior mode where many are zero and where most of their posterior mass is located near zero. A key advantage of the Bayesian LASSO formulation is that it can be made easily tractable via data augmentation. Park and Casella (2008) showed LASSO penalty corresponds exactly to the posterior mode if one put a double exponential (Laplace) prior on the β_j: $p(\beta_j) = \lambda/2 \exp(-\lambda|\beta_j|)$.

[10] The fix is adding a ridge penalty to make the data matrix invertible; however, the results seem rather sensitive to exactly the amount of ridge stabilization added and thus adds in yet another tuning parameter to the model. Appendix A outlines a more plausible way of choosing a ridge prior if this strategy is to be employed

[11] An advantage of the generalized LASSO algorithm, however, is that it automatically paths of the coefficients as λ grows. Appendix D discusses the way to replicate this behavior in this framework.

To make this tractable, they rely on a crucial result by Andrews and Mallows (1974):

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

This shows that we can represent the double exponential prior as a mixture of normal priors. Many other types of sparsity can be created by changing the mixing distribution on τ_j^2 (Polson and Scott 2011a).

$$\beta_j \sim N(0, \tau_j^2); \quad \tau_j^2 \sim \text{Exp}(\lambda^2/2) \tag{3}$$

If one were to take the joint density of $p(\beta_j, \tau_j^2)$ and integrate away the τ_j^2, one would exactly recover the double exponential prior that corresponds to the LASSO penalty. The posterior associated with the Bayesian LASSO is not inherently sparse, but a wide array of procedures exist for enforcing sparsity ex post (e.g. thresholding small coefficients, see Belloni and Chernozhukov 2013 for a discussion of the non-Bayesian case) or one can simply use the posterior itself. Kyung et al. (2010) extend these results to show how some other types of sparsity can be encoded.

To extend the vanilla case to the generalized case, however, some more work is needed. Starting with a simple case, the temporal LASSO involves specifying a dependent prior that is a simple location shift of the double exponential prior:

$$p(\beta_{j,t}|\beta_{j,t-1}) = \lambda/2 \exp(-\lambda|\beta_{j,t} - \beta_{j,t-1}|) \tag{4}$$

One can think of this as putting a vanilla sparsity penalty on the 'gap' between two coefficients in a time series; thus, at the optimum, it will shrink the gaps between two temporally adjacent coefficients to zero. Thus, upon augmentation, it becomes:

$$\beta_{j,t}|\tau_{j,t}^2, \beta_{j,t-1} \sim N(\beta_{j,t-1}, \tau_{j,t}^2) \tag{5a}$$

$$\tau_{j,t}^2 \sim \text{Exp}(\lambda^2/2) \tag{5b}$$

Again, if one were to integrate away the $\tau_{j,t}^2$, we recover a prior on $\beta_{j,t}|\beta_{j,t-1}$ that implies a posterior mode that exactly agrees with the temporal LASSO. This can be put even more generally to capture the types of sparsity listed in Table 1. Encoding the most generic form of the generalized LASSO as a Bayesian prior would suggest that the (marginal) prior on the coefficients

would be written as follows, where $\boldsymbol{D}$ has dimensionality $K \times p$. For notation, $w_{\boldsymbol{D}}$ is a constant that depends on $\boldsymbol{D}$ and the density is only proportional to a constant that does not depend on λ.

$$p(\boldsymbol{\beta}) \propto \lambda^{rank(\boldsymbol{D})} w_{\boldsymbol{D}} \exp(-\lambda ||\boldsymbol{D}\boldsymbol{\beta}||) \tag{6}$$

Appendix B derives this result—that is non-trivial—and shows that the prior corresponding to a generalized LASSO penalty and that enforces arbitrary types of complex sparsity for any $\boldsymbol{D}$ is written as above. It also discusses the conditions under which the prior is proper and how it can be modified to ensure this for rank deficient.[12] If $rank(\boldsymbol{D}) = p$, then the density above is normalized. In the case of $rank(\boldsymbol{D}) \neq p$, the Appendix shows that one can recast the problem into a generalized LASSO penalty on some m-dimensional vector of coefficients on an orthogonally rotated design matrix and thus derive the normalizing constant—although for inference, it is easier to work with $\boldsymbol{D}$. The joint (augmented) prior can be written as follows:

$$p(\boldsymbol{\beta}, \{\tau_k^2\}_{k=1}^K) \propto w_{\boldsymbol{D}} \lambda^{rank(\boldsymbol{D})-K} \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi\tau_k^2}} \exp\left(-\frac{\boldsymbol{\beta}^T \boldsymbol{d}_k \boldsymbol{d}_k^T \boldsymbol{\beta}}{2\tau_k^2} - \lambda^2/2\tau_k^2\right) \lambda^2/2 \tag{7}$$

This joint prior has some pleasing properties; first, the structure of the conditional priors, i.e. $\boldsymbol{\beta}|\{\tau_k^2\}_{k=1}^K$ and $\tau_k^2|\boldsymbol{\beta}$, have exactly the same form as in the vanilla case—normal and inverse Gaussian, respectively. The one caveat is that, for the generic case, the normal distribution that corresponds to $\boldsymbol{\beta}|\{\tau_k^2\}$ conditional prior may not be full rank. As shown in Appendix B, however, this can be cast as a singular multivariate normal—one generated from a full rank multivariate normal in m-dimensional—and thus the density can be represented using the Moore-Penrose pseudo-inverse, denoted by $\boldsymbol{A}^+$. When it is full rank, this equals the usual inverse. Noting that $\boldsymbol{\tau}$ stacks the $\{\tau_k^2\}$ into a diagonal matrix, the conditionals of the prior can be written as:

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

$$\tau_k^2|\boldsymbol{\beta} \sim \text{InverseGaussian}\left(\frac{\lambda}{|\boldsymbol{d}_k^T\boldsymbol{\beta}|}, \lambda^2\right) \tag{8b}$$

Thus, most of the existing architecture can be adapted for Bayesian inference in the generalized case. Appendix B outlines the particular $\boldsymbol{D}$ needed for each of the subtypes in Table 1, following

[12] A proper prior and thus proper posterior on $\boldsymbol{\beta}$ can be always guaranteed by either (a) adding a weak (but proper) normal prior in addition to the complex sparsity or (b) adding a vanilla LASSO penalization to all elements of $\boldsymbol{\beta}$. The first would correspond to a generalization of the "elastic net" (Zou and Hastie 2005) for complex sparsity. See Appendix B for a discussion of some of the subtleties of the elastic net analogue.

results from Tibshirani and Taylor (2011). This extends the existing work on Bayesian LASSOs using certain types of complex sparsity to include the existing cases outlined by other scholars (Kyung et al. 2010; Betancourt, Rodríguez, and Boyd 2017; Roualdes 2015; Faulkner and Minin 2018), but extends this work by allowing a general framework for the generalized LASSO to be integrated into the existing Bayesian methodology.[13] Further, the application to "categorical" or "geographic" sparsity, as discussed in the application, represents a novel extension of the generalized LASSO into the Bayesian domain (though see Pauger and Wagner 2018 who use a different regularization scheme, spike-and-slab, for Bayesian categorical fusion). Finally, this formulation sheds some interesting light on the the properties of a Bayesian generalized LASSO: Note that only when $rank(\boldsymbol{D}) = K$, i.e. it has fewer rows than columns such as in the case of functional or temporal sparsity, can the marginal prior on $\{\tau_k^2\}$ be written as a product of independent exponentials as in the vanilla case. Interestingly, for more complex types of sparsity, the marginal prior distribution of $\{\tau_k^2\}$ is rather complex; fortunately, we do not need to work with this directly for most of the problems discussed in this paper.

4 Fast Estimation of Complex Sparsity Using the EM Algorithm

Another major methodological contribution of this paper, however, is to leverage this integrated Bayesian representation to show that we can use an exact EM algorithm to find a posterior mode. The EM inference approach has a number of desirable properties; first, it has guaranteed convergence to a (local) posterior mode. If the likelihood function is concave and the penalties are all convex, the entire posterior is concave and thus has a single (unique) posterior mode that the EM algorithm is *guaranteed* to locate. This approach can be estimated in a way that avoids the issues with the existing framework for complex LASSOs insofar as, after augmentation, it is possible to update an entire sequence of connected coefficients at once—but cycle through *groups* of connected coefficients—using a variety of fast techniques that exploit the sparsity structure but still guarantee convergence. The major downside of this EM approach is that its convergence may

[13] Appendix B describes how the group LASSO is integrated into this framework. Some limited work uses a Bayesian implementation of the generalized LASSO, see Roualdes (2015) and Faulkner and Minin (2018) for an application to the functional LASSO. Kyung et al. (2010) propose a Gibbs Sampler for the sparse temporal LASSO that, while valid, is inefficient for any LASSO with a simple temporal [fusion] penalty as it does not exploit the standard method for estimating dynamic linear models—forwards filtering, backwards sampling (Früwirth-Schnatter 1994) and thus the sampler outlined in this paper will likely perform better in terms of mixing even on the simple temporal case, see also Betancourt, Rodríguez, and Boyd (2017) for an implementation along these lines.

be slow, it does not return standard errors, and it may get stuck in a local mode if the posterior is multimodal. For the first problem, there is a large literature on how to speed up EM algorithms—with or without maintaining deterministic convergence (McLachlan and Krishnan 2008).[14] The second problem (lack of standard errors) is more complicated, but one could rely on a bootstrap approach or adapting some existing theoretical work (e.g. Ratkovic and Tingley 2017). In this paper, I rely on the "hybrid" idea of using the EM algorithm to identify the structure of sparsity and then run the corresponding maximum likelihood model on the transformed design matrix (Tibshirani et al. 2004); this is discussed in detail later. The third is often not a problem as the likelihood function is concave and the penalty is convex so a unique mode exists, but sometimes arises in more complex cases.

Both of these initial issues can be addressed, however, by using the fully Bayesian approach that samples from the posterior and thus gets measure of uncertainty (via credible intervals) and can move between local modes. This approach may be slow, however, given that the EM approach if fast and ends up in an area of high posterior density, starting the algorithm from that point should lead to rapid convergence by the Gibbs Sampler.

To do all of this, a crucial analytical result is to show that, conditional upon augmentation, we can write all of these penalties in the framework of a normal linear model with highly sparse data when employing complex sparsity. Even in otherwise complex cases, the sparse nature of the data combined with the common forms of sparsity means that specialized techniques for solving high-dimensional systems can be leveraged to estimate these complex cases. Appendix B provides details, but it is illustrative to focus on a particular example: To preview the case outlined below, I re-examine the results in Trounstine 2016's paper on segregation and the provision of public goods, but I allow the effect of time to vary by city to address unobserved heterogeneity. As I show below, there are serious issues in terms of power and inference to including 2,000 city-specific linear time trends directly. For inference, however, the data matrix corresponding to that series of year-by-city has a huge dimensionality (about 14,000 by 4,000) but is almost entirely sparse. Thus, actually using this matrix in practice, as long as one relies on existing specialized techniques and matrix identities, is very efficient (Rue 2001; Bhattacharya, Chakraborty, and Mallick 2016) Unlike existing software that struggles to estimate such a model, and is unable to calculate cluster or robust standard errors, careful exploitation of the sparse structure makes this fairly doable.

In my re-analysis, I impose a type of geographic sparsity (cities within states are encouraged to

[14] Neal and Hinton (1998) provide a number of interesting suggestions in terms of sub-sampling in very large cases.

be equal), and this implies that $\boldsymbol{D}^T\boldsymbol{D}$ is also highly sparse.[15] Thus, as the relevant least squares that must be solved ($\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{D}^T\boldsymbol{\tau}^{-1}\boldsymbol{D}$) can be done very efficiently. In some other special cases (functional sparsity, temporal sparsity), even more efficient solvers can be employed.[16]

Another major benefit of this integrated Bayesian framework is that data augmentation can be used for non-linear outcomes to immediately, at almost no computational cost per iteration, extend fast estimation of complex sparsity to most types of non-linear models (binomial, negative binomial, and multinomial). Unlike existing implementations of the LASSO for non-linear models that require approximations (e.g. Fan and Li 2001), this means that the point estimates from this model correspond *exactly* the associated (penalized) maximum likelihood and thus are more suitable for inference than approximate methods. The exact algorithms for the non-linear models are sketched in Appendices C and E and work by employing other data schemes outlined in more detail elsewhere (Polson, Scott, and Windle 2013, see Goplerud 2018a for an overview with applications to political science). Because of the fact that a Gibbs Sampler exists to estimate the entire model, this means that the EM algorithm falls out rather automatically for most models.

I sketch the algorithm in the simplified linear model: Assume that our data are generated $y_i \sim N(\boldsymbol{x}_i^T\boldsymbol{\beta}, 1)$ where σ^2 is assumed known and equal to one for exposition only. We place a sparsity inducing prior of $p(\boldsymbol{\beta}) \propto \exp(-\lambda||\boldsymbol{D}\boldsymbol{\beta}||_1)$. To estimate the model via the EM algorithm, we rely on the results shown above about the representation of the penalty as conditionally normal or inverse-Gaussian after augmentation. The Gibbs Sampler samples from the corresponding distribution rather than replacing it with the mean or mode.

- For some number of iterations t, do the following, where superscript $^{(t)}$ denotes the results from iteration t.
  - Update τ_k^2 as follows (E-Step):

$$[(\tau_k^2)^*]^{(t)} = \frac{\lambda}{\boldsymbol{d}_k^T\boldsymbol{\beta}^{(t-1)}}$$

  - Perform a modified least squares update for $\boldsymbol{\beta}$ (ridge regression; M-Step):

$$\boldsymbol{\beta}^{(t)} = \left(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{D}^T[\boldsymbol{\tau}^{-1}]^*\boldsymbol{D}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}; \quad [\boldsymbol{\tau}^{-1}]^*_{i,i} = [(\tau_k^2)^*]^{(t)}$$

[15] Categorical sparsity, unfortunately, results in a dense $\boldsymbol{D}^T\boldsymbol{D}$ which for some dimensionality is fine, but at some point becomes too costly to use. One could adjust my assumption of state-specific geographic sparsity by perhaps having neighboring cities in different states also be linked.

[16] Functional sparsity is very similar and requires the usual trick of turning an autoregressive model into a state-space model (West and Harrison 1997). Group sparsity involves a dynamic linear model over all coefficients in the group.

Iterating this repeatedly is guaranteed to increase the log-posterior and converge to the global maximum if the objective function is concave. Further, note that if we have many variables that are separately modelled with complex sparsity (e.g. geographic sparsity for counties; functional sparsity for age, etc.), then these can be updated separately in an ECM (Meng and Rubin 1993) but still have guarantees on convergence as the penalties are additively separable across blocks (Tseng 2001).The extension to non-linear models is derived complete in Appendix C but simply involves one additional step in the sampler.

5 Estimating the Strength of Regularization

Throughout the paper, I have been assuming that the strength of prior encouraging sparsity (λ) is held constant. λ is a crucial parameter for models with sparsity as it governs the extent to which the resulting model is sparse; one can also think of it as encoding the strength of our prior belief that there is sparsity (e.g. neighbors are equal) in the underlying data. As λ goes towards zero, the effect of the prior or penalty goes away, and the resulting model is not sparse (e.g. no geographic fusion). As λ grows to infinity, the model will eventually become one that is "fully" sparse (e.g. all coefficients are fused together). Thus, deciding what value of λ to use when reporting results is an important choice. Appendix D discusses some further details, but I provide an overview of the major approaches in this section with some recommendations for applied researchers.

First, a classical way of selecting λ is cross-validation. Described extensively elsewhere (e.g. Hastie, Tibshirani, and Friedman 2009), this procedure splits the data into multiple parts, fits the model with a wide range of λ on some of the parts and examines its performance on the held-out parts. The λ that has the best performance—on some metric—on the held-out part is thus often chosen as the λ. In practice, however, there is a general rule to select a λ somewhat larger than what cross-validation suggests.

There are some potential issues with this framework for the cases of complex sparsity; first, in some cases, the data structure is not amenable to cross-validation. In the examples I consider below, I use complex sparsity to better estimate models with complicated fixed effects; however, as many of these have small numbers of observations per fixed effect, cross-validation that is stratified on the fixed effect seems inappropriate, as it leads to very few units left in the remaining model. Second, it can be quite prohibitive in terms of time as one needs to refit the model for each sub-fold in the most common K-fold validation and this cost increases dramatically as there are multiple tuning parameters.

A different approach that is deeply related to cross-validation is the large literature that relies on various information criteria (e.g. AIC or BIC) to select a model. The core idea is straightforward; we 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. We thus need 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. Gelman, Hwang, and Vehtari (2014) provide a nice discussion in the Bayesian case.

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 vanilla 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; this justification would also work for non-linear models in this paper by appealing to the Polya-Gamma data augmentation representation discussed in Appendix C. Some researchers have proposed biased-corrected measures of the degrees of freedom for the vanilla 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. 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, it agrees with the measure conjectured in Gertheiss and Tutz (2010) as the number of distinct fused 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.

The above discussion is focused on the non-Bayesian paradigm. In the Bayesian case, the question becomes what prior to give λ. While typically we want to give parameters weakly informative priors, this is not always desirable in the sparse case. If one gives a weak prior that does not grow with the data, then as the data grows, the prior will be dominated by the data and the resulting model will lose its sparse properties. Thus, general practice is to encourage the prior to grow with the data in some way. This allows the exact value of λ to adapt to the data, but it will be "anchored" in some sense around a theoretically desirable values. Again, there is a large literature on this, see Bühlmann and van de Geer (2011) for the canonical reference and Ratkovic and Tingley (2017) for a recent application in political science along this vein. Following existing research that suggests $\sqrt{N \ln P}$ in the vanilla case, I suggest that $\sqrt{N \ln rank(\boldsymbol{D})}$ is plausible for the generalized case, although this needs further investigation.

Overall, therefore, I have outlined two strategies for tuning the strength of regularization.[17] First, if one is relying on the non-Bayesian paradigm, then one should use either cross-validation or an information criteria (AIC, BIC, etc.) to select the value of λ. However, an important robustness check would be to show the evolution of the quantity of interest (e.g. the treatment effect) over a range of λ while noting which points correspond to the optimal ones suggested by different metrics. If the quantity of interest is robust to the choice of λ in plausible ranges, that should give confidence in the conclusions from using complex sparsity. Second, if one is relying on the fully Bayesian paradigm, one should anchor the prior mean of λ at $\sqrt{N \ln rank(\boldsymbol{D})}$.

[17] A third, albeit more non-standard, method is discussed in Appendix D. This approach puts λ into the EM algorithm directly. The simplest, but conceptually odd, puts it as a parameter to be maximized in the M-Step, i.e. finding the posterior mode of $p(\boldsymbol{\beta}, \lambda | \boldsymbol{y}, \boldsymbol{X})$ instead of the usual $p(\boldsymbol{\beta} | \lambda, \boldsymbol{y}, \boldsymbol{X})$. While used in some research, e.g. Polson and Scott (2011b), Ratkovic and Tingley (2017), and Goplerud et al. (2018), it has some difficulties that are discussed in the Appendix. Alternatively, one can put λ into the E-Step and target the correct $p(\boldsymbol{\beta} | \boldsymbol{y}, \boldsymbol{X})$ at the cost of needing to do a small number of one-dimensional numerical integrals to evaluate the moments of λ.

6 Extension to Other Forms of Regularization

In the above discussion, I have used penalties that are defined by an absolute value penalty (ℓ_1-norm), i.e. $\lambda|\beta_j|$. However, as is widely known, this penalty suffer from a number of problems; first, and most importantly, it severely biases (attenuates towards zero) the non-zero coefficients that are recovered from the model. Ideally, we would have a form of regularization that achieves some desirable properties known as the 'oracle' properties: Roughly, as the amount of data grows, our estimator should (i) give non-zero values only to the *true* non-zero coefficients and (ii) be unbiased for those non-zero coefficients.[18] A large literature has tried to address the problems with the LASSO in the vanilla case, Polson and Scott (2011a) provides a good overview. However, for the more complex cases (e.g. temporal sparsity), almost all work in the 'point estimation' case and most Bayesian analysis relies on the (standard) LASSO and thus inherits those undesirable properties. The framework outlined in this paper allows for many modern types of regularization to be applied to *any* complex sparsity thus immediately expanding the repertoire available to social scientists. Many modern forms of regularization (sometimes known as global-local priors or continuous sparse signals) can be cast in the following hierarchical form for some distribution g:

$$\beta_j|\tau_j^2 \sim N(0, \tau_j^2); \quad \tau_j^2|\lambda \sim g \tag{9}$$

As outlined above, the LASSO sets g to be $\text{Exp}(\lambda^2/2)$. Following Polson and Scott (2011a), these models are typically known as global-local shrinkage models as there are some shared parameters (λ - the 'global') that control the global level of sparsity, but each variable has its own specific augmentation variable (τ_j^2 - the 'local') that controls whether it, *in particular*, is shrunk towards zero. By carefully selecting the g function, i.e. how the local variable is specified, one can create a huge array of models that have many different and desirable properties. For example, if τ_j^2 is given a half-Cauchy prior, this generates the horseshoe penalty (Carvalho, Polson, and Scott 2010). Importantly for inference, this global-local prior means that, conditional on the τ_j^2, one is simply imposing a normal prior on the coefficients and thus can rely on standard forms of inference where β_j has a normal prior. For carefully chosen g, the conditional distribution of $\tau_j^2|\boldsymbol{\beta}$ is also tractable. An extensive review can be found in Polson and Scott (2011a).

In the case of complex sparsity, the issue becomes somewhat more complex. In the case of

[18]The oracle property is phrased in terms of the vanilla LASSO, e.g. Zou (2006), as having property (i) and then property (ii) as converging to the maximum likelihood estimator for the model estimated only on the truly non-zero coefficients. As the analogue to the maximum likelihood estimator is trickier to conceptualize in the temporal and complex case, the phrasing of (ii) above captures the key point—unbiasedness.

$rank(\boldsymbol{D}) = K$, as shown above, we can cast the LASSO-based sparsity as having independent exponential distributions on τ_k^2. By switching that g distribution to something else, we directly generalize that type of complex sparsity to other global-local priors. In the other cases, the same issue as above arises—the marginal distributions on the τ_k^2 no longer have a clean, independent form. However, that may not, in fact, impact our ability to engage inference as long as we specify the joint prior appropriately:

$$p(\boldsymbol{\beta}, \{\tau_k^2\}_{k=1}^K) \propto \exp\left(-\frac{\boldsymbol{\beta}^T \boldsymbol{D}^T \boldsymbol{\tau}^{-1} \boldsymbol{D}\boldsymbol{\beta}}{2}\right) \prod_{k=1}^K \frac{g(\tau_k^2)}{\sqrt{\tau_k^2}} \tag{10}$$

This is prior is not normalized, and thus calculating and evaluating the normalizing constants insofar as they depend on other hyper-parameters may be challenging, but sampling from this prior—and thus the posterior on $\boldsymbol{\beta}$ and τ_k^2 should be tractable inasmuch as Bayesian inference on the vanilla case is tractable conditional on those hyper-parameters.

In the empirical applications, I focus on one specific and highly common extension in the non-Bayesian case. In attempting to improve the vanilla LASSO, Zou (2006) defines the 'adaptive LASSO' as follows: The penalty for each variable is scaled by some weight w_j^γ, i.e. $p(\beta_j) \propto \exp(-\lambda/w_j^\gamma|\beta_j|)$. In that paper, w_j is defined as 'some $\sqrt{N}$-consistent estimator' (e.g. the maximum likelihood estimate). The idea is that the MLE for the entire model is not a terrible guess for the value and thus if the MLE is small, the true value of the coefficient should be small so a small MLE results in a *big* shrinkage. Zou (2006) shows that, assuming λ grows at the appropriate rate, this estimator has the oracle property. A large, but rather disconnected, literature has shown that various types of complex sparsity can also have the oracle property given similarly straightforward adaptive weights.[19] The cost of the adaptive LASSO is that it adds an additional regularization term (γ) is rather important for small-sample performance and must be calibrated via the same set of strategies noted above for tuning λ; the most common default is to set $\gamma = 1$. In terms of the global-local representation of the (vanilla) Bayesian adaptive LASSO,

$$\beta_k \sim N(0, \tau_k^2); \quad \tau_k^2 \sim \text{Exp}\left(\frac{\lambda^2}{2w_k^{2\gamma}}\right) \tag{11}$$

As noted above and discussed in detail in Appendix E, this can be adapted to the generalized case for arbitrary $\boldsymbol{D}$. Inference on λ proceeds analogously to the standard LASSO case, although weighted by the adaptive weights, but inference on γ is tricky for a general $\boldsymbol{D}$. Thus, one can

[19]Most proof strategies target a specific type of sparsity, e.g. temporal, and thus consulting the canonical papers typical comes with an oracle result for that case. Working on a general proof for oracle property for the generalized LASSO is an interesting case for future research.

simply assume it to be fixed at one, as is standard, or rely on more complicated methods to sample γ to address the unknown normalizing constant that depends on this parameter.

7 Incorporating Random Effects

Another key tool in the social scientist's arsenal are random effects, i.e. some shared 'shock' or effect between groups. That is each observation i has some vector of covariates and/or membership dummy variables $\boldsymbol{z}_i$ that is multiplied by a vector of coefficients $\boldsymbol{\alpha}$ that is given a normal prior—that is, generally, not sparsified.[20] These can also be added to the model; as this has been discussed more extensively elsewhere (Goplerud 2018a), I defer discussion of this to those papers. In short, it requires a mean-field variational approximation to the E-Step of the EM algorithm; it adds no complication to the Gibbs Sampler.

8 Applying Complex Sparsity to Fixed Effects

I focus on one important application of complex sparsity—estimating fixed effects in a more flexible way. Consider the scenario outlined at the beginning of the paper: A recent influential paper by Trounstine 2016 examines how residential segregation in American cities affects the provision of public goods and political outcomes. The paper engages in a number of different analyses that employ fixed effects that I re-analyze here in the context of complex sparsity.

The first analysis (Table 1; Trounstine 2016, p. 713) gathered extensive political data on 25 cities across a number of years (91 observations in total) to analyze how residential segregation affects the racial divide at elections. The outcome variable is the largest pairwise difference between racial groups (white vs black; white vs Latino; black vs. Latino) for the winning candidate at the relevant mayoral election. The main explanatory variable is residential segregation measured by Theil's H index; roughly, it "measures the degree to which the diversity in each neighborhood differs from the diversity of the city as a whole" (Trounstine 2016, p. 711). Additional controls are included; these are outlined in Appendix F and described at length in the original paper. To control for possible unobserved heterogeneity, the researcher has a number of possible strategies at their disposal of which fixed effects is perhaps the most popular. Trounstine 2016 follows in this tradition and includes two-way fixed effects. One, for year, captures temporal variation in

[20] There is some work on inducing sparsity in the random effects directly (Kinney and Dunson 2007; Bondell, Krishna, and Ghosh 2010; Ibrahim et al. 2011); as this is not the key point of this project, I do not explore that in detail but this could be added to the model without much additional difficulty.

the data. The other is designed to capture spatial heterogeneity, i.e. differences between cities. However, unlike the usual approach, the author includes *region* (not city) fixed effects and city *random* effects. The author's decision to note include city fixed effects is not elaborated in detail in the main text, but presumably is because four cities occur for only one year and there is some perfect collinearity with those singleton cities as other variables in the analysis (e.g. the only election for Austin is also the only election for the year 2009).

While this was a reasonable choice by the author, it is rather sensitive to a number of choices. First, if one estimates the random effects via REML (the default in `lmer` in `R`; Bates et al. 2015), the main coefficient is no longer statistically significant at the 0.05 level (p-value 0.064)—whereas it is significant if one uses maximum likelihood (the default option in the STATA replication code). Second, and more importantly, if one actually examines the random effects—estimated by either method, we see that, in fact, they are set to zero as the variance of the random effects themselves is effectively zero. Thus, there is effectively *no* control for city-level heterogeneity in the model as a model estimated with OLS, excluding the random effects, has identical point estimates to the random effects model.

Most importantly, it is actually possible to use fixed effects for cities as long as we are willing to accept a smaller number of observations: By excluding those four observations above, there is sufficient variation in enough cities to estimate all of the coefficients in the original model. This specification is the most causally robust—it leverages within city variation to estimate the effect of residential segregation. However, given the small number of observations, we have an issue of power: Running this model gives a much larger point estimate (3.841 vs 0.932), but the standard error is about ten times as large (p-value of 0.40). This leaves an open question of the robustness of the result in the original paper. The existing specification provides effectively no control for city level heterogeneity but a result that does control for this in the standard way is significant, if seemingly highly under-powered.

Complex sparsity provides a way to bridge the gap between the author's intention—control for unobserved heterogeneity at the city level—and the practicalities of the data. The intuition is straightforward: Imagine it was known that certain cites shared the same unobserved heterogeneity, i.e. they had identical fixed effects. In that case, we would simply want to include dummies for those *groups* to control for the unobserved heterogeneity. Insofar as those groups had variation on all of the variables of interest, then the model is identified. As these groups contains multiple cities and thus more variation, this will have more power than the simple city-fixed effects appro-

ach. It also will allow for city-level heterogeneity to remain in the model, unlike in the random effects case.

Of course, the major snag in this thought experiment is that the groups are unknown. Thus, one plausible strategy is to estimate them. One existing approach ("grouped fixed effects"; Bonhomme and Manresa 2015) does this by, effectively, clustering the residuals of a linear regression using k-means and using that to assign group membership by iterating between k-means and regressions including those groups. The authors show that under a variety of regularity conditions, their procedure will asymptotically recover the correct groups and point estimates on the main coefficients of interest. This procedure, however, does not clearly correspond to a probabilistic model and thus many extensions outlined in this paper, such as extensions to non-linear models as well as valid measures of uncertainty, are difficult to include. Further, the types of groups permitted by that model are somewhat limited, whereas complex sparsity easily allows, for example, for groupings that respect network structure (e.g. geography), different groupings across different variables, etc.

The other major approach comes from frequentist applications of a specific application of complex sparsity (Categorical Sparsity). Gertheiss and Tutz (2010), but see Bondell and Reich (2009) for some earlier work on the ANOVA case, derived an estimator corresponding to the posterior mode of a linear model with categorical sparsity on some variable of interest (city, in our example). At the optimum, this will fuse together many levels of state-year into groups exactly as in the group-fixed effects framework. They rely on the adaptive LASSO penalty as discussed above, with a modification to account for the different sizes of the groups in the data. They show, and Chen (2015) shows in the logistic case, that under the usual adaptive LASSO conditions—with some additional regularity conditions concerning the non-trivial asymptotic size of each group–their model has the oracle property for recovering the groups and unbiased estimation of the (true) non-zero differences between groups. The work in this paradigm is frequentist, and thus this paper appears to be the first generalization of the generalized LASSO into the Bayesian framework, although other approaches for collapsing categorical predictors in a fully Bayesian framework remains an active area of research (Pauger and Wagner 2018).

One can think of the complex sparsity approach as a bridge between the fixed effects and random effects traditions (Tutz and Oelker 2017); while the usual random effects tradition places a normal prior on the city effects, complex sparsity places a prior that allows for dependence between city dummies and induces grouping in the posterior mode. This interpretation also explains why

we can include variables that are perfectly collinear with city in the analysis (as is possible in the classical random effects tradition), but also gives us the power of the interpretation of the fixed effects as 'de-meaning' asymptotically as the asymptotic behavior of the oracle complex sparsity is *not* for each fixed effect to converge to unique value for each city but rather that the true *groups* emerge and thus the results can be interpreted in the usual "within" fixed-effects context conditional on the group. Thus, in the specification that includes all cities (including those only observed for a single year) or variables that are perfectly correlated with city (e.g. non-partisan primary), the model is only identified by virtue of the prior, as in the case for random effects. However, unlike random effects, the asymptotic behavior—assuming λ grows at the appropriate rate—is to identify a model that *is* identified.

For now, I focus on the frequentist approach to illustrate the properties of complex sparsity. To begin, I re-specified their model using the adaptive LASSO.[21] I considered two possible patterns of sparsity; first, I use categorical sparsity to fuse together any possible combinations of city effects.

I proceed as follows: I vary λ from very small (little fusion into groups) to large (all city-effects fused into one group) and note which λ would be selected by the AIC and BIC criteria discussed above. Following existing practice, I re-fit the model on the groups that are selected using ordinary least squares; this is the "hybrid" approach of Tibshirani et al. 2004, see also relatively common practice in applied research (e.g. Gertheiss and Tutz 2010; Hui, Warton, and Foster 2015) and has some theoretical underpinning in the case of the vanilla LASSO (e.g. Belloni and Chernozhukov 2013). A benefit of this approach is that we can easily get standard errors (robust or otherwise) from this procedure.[22]

To begin, Figure 1 shows how complex sparsity affects the structure of the unobserved heterogeneity. The left panel shows the estimated coefficients using complex sparsity as λ varies; we see that as λ increases, more and more of the city effects are fused together until, eventually, all are set to zero. The right panel shows the number of groups as the regularization strength varies, calculated by assuming that all city effects that are closer than 0.0001 should be fused together. Both figures have a vertical red line to indicate the optimal model as selected by the AIC and BIC—in this case, the two measures agree.

The primary quantity of interest in this analysis, however, is the effect of residential segregation.

[21] I fix $\gamma = 1$; note that even though the model is un-identified, only the differences between the city coefficients matters for the adaptive weights; this quantity is identified. I put a diffuse ridge prior on the variables to stabilize the adaptive weights.

[22] An important future extension would adapt the sandwich estimator from Zou (2006) to the complex sparsity case. This is likely possible, see Gertheiss and Tutz (2010) and Chen (2015), who derive the asymptotic normality results for categorical sparsity.

Figure 1: Effect of Complex Sparsity on City-Level Heterogeneity

(a) Regularization Path

(b) Number of Unique Groups

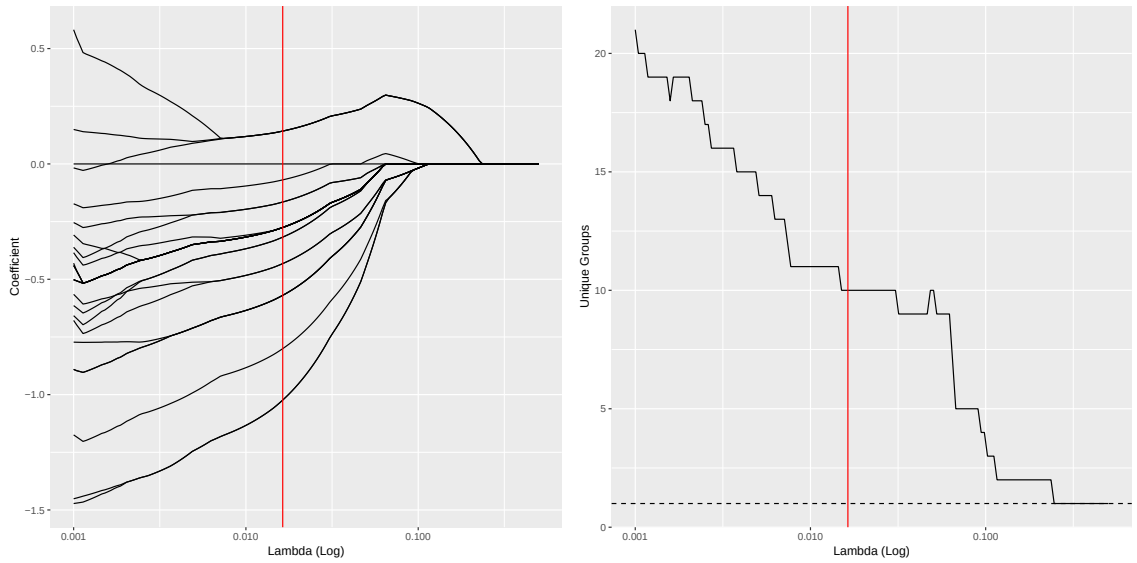


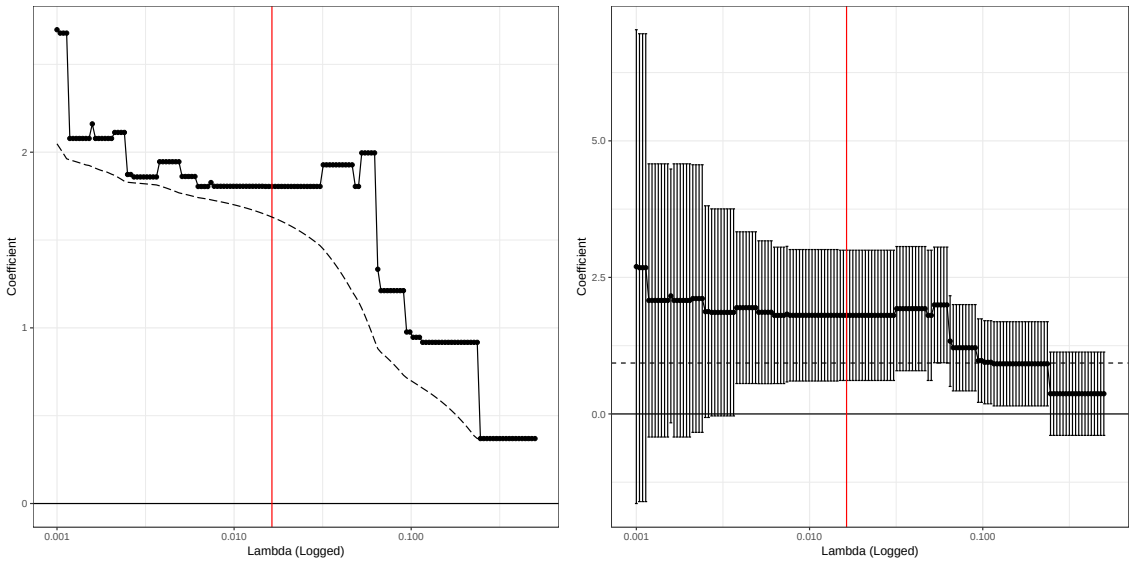
Figure 2 shows the results for that variable. The left panel shows in a dashed line the point estimates from the model with the adaptive penalty. The solid line with dots indicates the point estimates from the refit model, i.e. linear regression run when the groups are assumed known (i.e. when the coefficients are closer than 0.0001). We see that they are fairly close suggesting that the attention from using the adaptive LASSO is rather small in this instance. The right panel adds confidence intervals to the refit model to examine the statistical significance of this refit model. We see that, for almost all values of λ, there is a statistically significant effect as soon as we impose a small amount of heterogeneity to structure the city-specific effects. The dashed line indicates the original point estimates in Trounstine 2016 and we see that our point estimate is larger than the results in the original paper, although less larger than in the pure city-fixed effects case. The groups between the cities allows us to gain efficiency, however, as the result is clearly statistically significant.[23]

Overall, therefore, complex sparsity allows us to salvage the initial results of Trounstine 2016 in a more causally robust way; rather than believing the reported analysis that does nothing to control for city-level heterogeneity or finding a null effect by an under-powered analysis that exploits only within-city variation, complex sparsity lets us say, if we are willing to make an assumption that

[23]The results here are calculated using regular standard errors, as in Trounstine 2016. Calculations with clustered standard errors on either city or group return a similar result; indeed, they are in fact smaller than the regular standard errors.

Figure 2: Effect of Complex Sparsity on Segregation Index

(a) Penalized versus Refit Coefficient Path (b) Refit Path with Standard Errors



there is some underlying grouping of the heterogeneity *across* cities, we can leverage that to gain more power when drawing inferences. Indeed, as we see that for most levels of sparsity that are imposed, i.e. the strength of our belief that there is grouping between cities, we find a significant effect for residential segregation on the racial divide in mayoral elections. Thus, in cases where fixed effects strategies are infeasible or underpowered, and random effects fail to adequately control for the underlying heterogeneity, complex sparsity provides a path forwards.

9 Using Complex Sparsity to Estimate Time Trends

Another way to apply complex sparsity to the question of fixed effects involves the question of time trends. A well-known threat to confounding for fixed effects or difference-in-differences designs is differential time-trends between the units. If we knew that the time-varying confounding had a linear trend with respect to time, one could include unit-specific time-trends (of some order) to address this. However, this comes a heavy cost: It adds a huge number of parameters to the model and, for models with fairly short time panels, may severely under-power the resulting estimates. Further, fitting these models can be computationally intensive as one is estimating two-times the number of units parameters as nuisance fixed effects.

For a concrete application, consider the second series of results in Trounstine 2016. The

previous analysis was a small-data context where high-quality data existed on political outcomes; the second analysis uses a large-n framework where data on approximately 2,000 cites surveyed at six time periods (1982 to 2007, every five years) giving us around 14,000 observations. In this analysis, Trounstine 2016 examines the role of segregation on spending on public goods to compliment the analysis of political outcomes. For this analysis, the outcome variable is the "direct general expenditure" per capita. To address unobserved confounding, the author employs city-level fixed effects to leverage within-city variation to more credibly estimate the causal effect. However, the analysis does not take into account the role of *time*, as there are no controls for year. As spending may have secularly increased, or decreased, over time, and this may be correlated with variable of interest, this poses a threat to identification that is worth carefully examining.

To illustrate the potential for concern, Table 2 reports the main effects using a number of specifications that control for time. First, Model 1 is the result in (Trounstine 2016, p. 715). Second, Model 2 is the result controlling for time linearly and Model 3 is the result controlling for time using dummy variables for each year.

Table 2: Controlling for Time

Statistic	(1)	(2)	(3)	(4)	(5)	(6)
Coefficient	-1.131	-0.460	-0.560	-0.637	-0.204	-0.471
Standard Error	0.172	0.178	0.179	0.330	0.183	0.115
t-stat	-6.585	-2.585	-3.125	-1.931	-1.114	-4.082

Note: Model 1 replicates the results in Trounstine (2016, p. 715), Table 2. Model 2 includes a linear control for year; Model 3 includes year fixed effects. Model 4 includes city-specific time trends. Model 5 includes state-specific time trends. Model 6 is the AIC/BIC selected model using geographic sparisty described in the main text. For comparision across models, all include regular, i.e. non-robust, standard errors. The standard error on residential segregation increases if robust standard errors are used for Models 1-3.

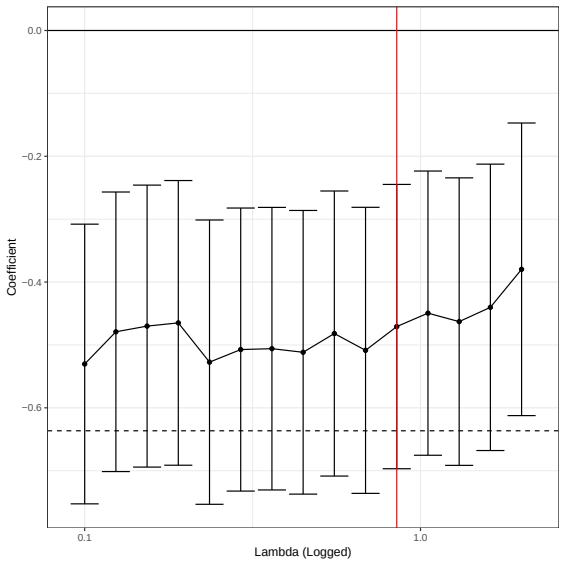
We see that time seems to matter in the aggregate; the main effects are noticeably diminished, although still significant and non-trivial in magnitude. Yet, this robustness test assume thats time has the same impact across all units; if it were the case that certain influential units had differential time-trends that were correlated with their trends in segregation, this may further affect the robustness of the results. One could test this in a number of ways; first, the most robust would be to include city-specific time trends at the cost of a large number of degrees of freedom (Model 4 in Table 2). One could try a more limited test of allowing state-specific time trends (Model 5 in Table 2). We see that the point estimates are broadly comparable (albeit smaller now that time is included), but the standard errors have grown enormously—even using regular and not clustered standard errors that, in fact, cannot be computed using standard software (even after de-meaning) because of the dimensionality of the dataset. Thus, it is hard to know whether

the null result that accompanies these point estimates is due to actual confounding by time or a simply underpowered model.

Again, I turn to complex sparsity to help address the issue. Using a very similar procedure to above, I run a model that includes (un-regularized) fixed effects for every city and the additional controls included in the original paper.[24] To deal with the time-varying confounding, however, I rely on geographic sparsity. Given that there are over 2,000 cites, it is computationally infeasible on a standard computer to allow every city to be equal to every other; that would imply over 1.9 million rows in $\boldsymbol{D}$! Thus, I use the my theoretical knowledge to assume that it is plausible that cities in a *state* may share a common time trend. The fully sparse version of this model (constant time trends inside a state) is reported as Model 5 in Table 2.[25]

Figure 3 shows, analogously to Figure 2, the path of the coefficient on segregation as λ varies. The dashed line includes the original point estimates from using city-specific time-trends (Model 4). We see that while the point estimate is smaller, the standard errors are also smaller and thus the effect remains significant. The model selected by the AIC and BIC (that again agree) contains about 300 distinct groups that each have their own random slope for time. While more than the fifty-some parameters added if we assumed the time trend was constant within states, this is also many fewer than the 2,000 city-specific slopes that the naive fixed effects approach would include.

Figure 3: City-Specific Time Trends with Geographic Sparsity



[24] See Appendix F for details on the controls.

[25] I exclude states with only one city (Alaska, Hawaii, and Washington DC) for the sake of simplicity. Those cities could be fused to neighboring states, perhaps. I also explicitly exclude the cities with only one election as these cannot contribute to estimated coefficients when city fixed-effects are included.

10 Conclusion

This paper began by noting that many of functional form assumptions in our models can be thought of as encoding particular types of sparsity. When one thinks that variables are stable across time and space, this is temporal and geographic sparsity. When one thinks that a variable has a linear effect, this is 'functional' sparsity. And so on. The goal of this paper is to show that extensions of common methods in machine learning are designed *exactly* to address these problems that are endemic in social sciences. By deriving a framework for stable, fast, and tractable estimation of a vast array of complex types of sparsity, this paper makes testing those assumptions within reach of the applied researcher.

Rather than having to rely on specialized inference procedures that are not suited for general social scientific research,[26] I derived and presented a general framework for estimating complex types of sparsity that is fast and works on many major dependent variables encountered in typical research (linear, binary, multinomial, and negative binomial). The key innovation was to cast the complex sparsity in a Bayesian fashion and then note that the EM algorithm allows us to estimate this in a rapid and stable way—even for non-linear outcomes—as well as adding other common features like random effects. Further, to get a model that has good performance, I extend the existing work on complex sparsity to enable the use of other sparsity-inducing penalties that is known to have much better theoretical properties than the LASSO in a general case. Thus, the framework outlined here would improve existing applications of complex sparsity; examining and demonstrating that is a fruitful area for future research.

I then applied this framework to an important recent paper (Trounstine 2016) in American politics that examines the role of residential segregation on public goods provisions and political outcomes. I find that if one applied a number of standard "robustness tests" to the author's original findings, the results go away—that is, become insignificant. In the case of political outcomes, a stricter test using city-specific random effects (vs. the existing model that only controls for region-level heterogeneity) increases the size of the coefficient but explodes the standard error. In the case of public goods, controlling for time in the most robust way (including city or state specific linear time trends) also renders the main effect insignificant and noticeably decreases its magnitude. Given that assuming some structure across cities is reasonable, this suggests that the results *are* robust to possible unobserved confounding, although tests without using complex sparsity would

[26] For example, the most common existing packages either assume a single variable, that all variables are sparsified in the same way, or make one painstakingly enter the types of sparsity via specifying $\boldsymbol{D}$. They are also not able to estimate non-linear models in the generalized case.

have suggested otherwise.

It is worth stressing that the patterns of heterogeneity uncovered by complex sparsity were not something that the existing literature had strong theoretical guidance behind. It would be hard to justify and defend to oneself and others that specific groups of cities should or shouldn't be clustered together for fixed effects or linear time trends. Indeed, whether there should be any clustering at all is open to debate and can be determined by examining criteria like the AIC or BIC. Complex sparsity allows researchers to address that question by saying that if we are willing to put some belief on there being structure in the heterogeneity (i.e. groups of units with similar effects in the example in this paper), we have a principled procedure to identify those groups without resorting to many of the issues associated with "p-hacking". By showing the results across a range of possibility strengths of prior belief on the structure of the groups, as well as noting which level is best supported by the data using an information criterion-based approach, researchers can transparently show to readers how much prior belief in sparsity matters for their conclusions.

Thus, the power of this framework is that it allows researchers to use theoretical assumptions or knowledge—that usually comes at the level of expecting certain types of geographic or temporal relationships between variable—without pre-judging exactly where, when, and how the heterogeneity should appear. This is the power of the entire enterprise of sparsity in the traditional sense; this paper gives the tools for leveraging that enterprise to tackling the core types of heterogeneity that appear in social scientific questions.

References

Ali, Alnur, and Ryan J. Tibshirani. 2018. "The Generalized LASSO and Uniqueness". *arvix preprint*. `https://arxiv.org/pdf/1805.07682.pdf`.

Andrew, Galen, and Jianfeng Gao. 2007. "Scalable Training of L_1-Regularized Log-Linear Models". In *International Conference on Machine Learning 2007*.

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.

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.

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

Belloni, Alexandre, and Chernozhukov. 2013. "Least Squares After Model Selection in High-Dimensional Sparse Models". *Bernoulli* 19 (2): 521–547.

Betancourt, Brenda, Abel Rodríguez, and Naomi Boyd. 2017. "Bayesian Fused Lasso Regression for Dynamic Binary Networks". *arvix preprint*. `https://arxiv.org/pdf/1710.01369.pdf`.

Bhattacharya, Anirban, Antik Chakraborty, and Bani K. Mallick. 2016. "Fast sampling with Gaussian scale mixture priors in high-dimensional regression". *Biometrika* 103 (4): 985–991.

Bondell, Howard D, Arun Krishna, and Sujit K Ghosh. 2010. "Joint Variable Selection for Fixed and Random Effects in Linear Mixed-Effects Models". *Biometrics* 66 (4): 1069–1077.

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.

Bühlmann, Peter, and Sara van de Geer. 2011. *Statistics for High-Dimensional Data*. Springer.

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

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`.

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

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

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

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.

— . 2008. "Sparse Inverse Covariance Estimation with the Graphical Lasso". *Biostatistics* 9 (3): 432–441.

Friedman, Jerome, et al. 2007. "Pathwise Coordinate Optimization". *The Annals of Appplied Statistics* 1 (2): 302–322.

Früwirth-Schnatter, Sylvia. 1994. "Data augmentation and dynamic linear models". *Journal of Time Series Analysis* 15 (2): 183–202.

Gelman, Andrew, Jessica Hwang, and Aki Vehtari. 2014. "Understanding predictive information criteria for Bayesian models". *Statistics and Computing* 24 (6): 997–1016.

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

Goplerud, Max. 2018a. "Power of Pólya-Gamma: Non-Linear Models Made Normal". *Working Paper*.

— . 2018b. "Selecting Interactions using Complex Sparsity". *Working Paper*.

Goplerud, Max, et al. 2018. "Sparse Multilevel Regression (and Poststratification [sMRP])". *Working Paper*. `https://scholar.harvard.edu/files/dtingley/files/sparsemultilevel.pdf`.

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

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.

Ibrahim, Joseph G, et al. 2011. "Fixed and random effects selection in mixed effects models". *Biometrics* 67 (2): 495–503.

Imai, Kosuke, and In Song Kim Kim. Forthcoming. "When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data?" *American Journal of Political Science*. `https://imai.fas.harvard.edu/research/FEmatch.html`.

Johndrow, James E., Paulo Orenstein, and Anirban Bhattacharya. 2018. "Scalable MCMC for Bayes Shrinkage Priors". *arvix preprint*. `https://arxiv.org/pdf/1705.00841.pdf`.

Kim, Seung-Jean, et al. 2009. "L-1 Trend Filtering". *SIAM Review* 51 (2): 339–360.

Kinney, Satkartar K, and David B Dunson. 2007. "Fixed and random effects selection in linear and logistic models". *Biometrics* 63 (3): 690–698.

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

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

Meng, Xiao-Li, and Donald B. Rubin. 1993. "Maximum likelihood estimation via the ECM algorithm: A general framework". *Biometrika* 80 (2): 267–278.

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.

Neal, Radford M., and Geoffrey E. Hinton. 1998. "A view of the EM algorithm that justifies incremental, sparse, and other variants." In *Learning in Graphical Models*, ed. by Michael I. Jordan, 355–368. Springer.

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, 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.

Ratkovic, Marc, and Dustin Tingley. 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.

Rue, Håvard. 2001. "Fast sampling of Gaussian Markov random fields". *Journal of the Royal Statistical Society. Series B (Methodological)* 63 (2): 325–338.

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

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, Robert, et al. 2005. "Sparsity and Smothness Via the Fused LASSO". *Journal of the Royal Statistical Society. Series B (Methodological)* 67 (1): 91–108.

Tibshirani, Ryan J. 2014. "Adaptive Piecewise Polynomial Estimation Via Trend Filtering". *The Annals of Statistics* 42 (1): 285–323.

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.

Trounstine, Jessica. 2016. "Segregation and Inequality in Public Goods". *American Journal of Political Science* 60 (3): 709–725.

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

Tutz, Gerhard, and Margret-Ruth Oelker. 2017. "Modelling Clustered Heterogeneity: Fixed Effects, Random Effects and Mixtures". *International Statistical Review* 85 (2): 204–227.

West, Mike, and Jeff Harrison. 1997. *Bayesian Forecasting and Dynamic Models*. 2nd. Springer.

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.

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

Zou, Hui, and Trevor Hastie. 2005. "Regularization and Variable Selection via the Elastic Net". *Journal of the Royal Statistical Society. Series B (Methodological)* 67 (2): 301–320.

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 Details on the Generalized LASSO Prior

This appendix focuses on deriving some results related to the prior corresponding to the generalized LASSO penalty. It proves that the normalizing constant in terms of λ can be drawn out as in the main text and discusses some issues that arise when employing other types of sparsity.

To begin, as noted above, the un-normalized density for the marginal generalized LASSO Bayesian prior can be written as follows:

$$p(\boldsymbol{\beta}) = c(\lambda, \boldsymbol{D}) \exp(-\lambda ||\boldsymbol{D}\boldsymbol{\beta}||_1) \tag{12}$$

For fully Bayesian inference on λ, it is important to know whether we can characterize the normalizing constant in terms of two separate terms $c(\lambda, \boldsymbol{D}) = k(\lambda) w_{\boldsymbol{D}}$. If so, then it is possible to accurately estimate a fully Bayesian model or EM algorithm where λ is a parameter. In doing so, some interesting facts about the generalized LASSO are derived. I start by decomposing $\boldsymbol{D}$—unlike Arnold and Tibshirani (2016) who rely on a QR decomposition, I use the singular value decomposition, noting that it can be written as follows, with the dimensionality of each matrix indicated in subscript. Assume the dimensionality of $\boldsymbol{D}$ is m. Recall that $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal and $\boldsymbol{\sigma}$ is a full-rank diagonal matrix. Further, note that $p \geq m$.

$$\boldsymbol{D}_{K\times p} = \boldsymbol{U}_{K\times K} \begin{pmatrix} \boldsymbol{\sigma}_{m\times m} & \mathbf{0}_{m\times p-m} \\ \mathbf{0}_{K-m\times m} & \mathbf{0}_{K-m\times p-m} \end{pmatrix}_{K\times p} [\boldsymbol{V}_{p\times p}]^T \tag{13}$$

We can simplify this further in some crucial ways; first, define $\tilde{\boldsymbol{U}}$ as the first m columns of $\boldsymbol{U}$. In the same vein as a 'reduced rank' QR factorization, we can simplify the presentation of $\boldsymbol{D}$.

$$\boldsymbol{D}_{K\times p} = \tilde{\boldsymbol{U}}_{K\times m} [\boldsymbol{\sigma}_{m\times m}, \quad \mathbf{0}_{m\times p-m}] [\boldsymbol{V}_{p\times p}]^T \tag{14}$$

Turning back to the prior, we define $\tilde{\boldsymbol{\beta}} = \boldsymbol{V}^T\boldsymbol{\beta}$. $\boldsymbol{V}$ is full rank and orthogonal so we can apply the change of variables formula. Note, this would be equivalent to re-specifying the model with a new data matrix $\tilde{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{V}$. After doing this, the prior would have the following density:

$$p(\tilde{\boldsymbol{\beta}}) = c(\lambda, \boldsymbol{D}) \exp(-\lambda ||[\tilde{\boldsymbol{U}}\boldsymbol{\sigma}, \quad \mathbf{0}]\tilde{\boldsymbol{\beta}}||) \tag{15}$$

This leads to the interesting observation that, in fact, only the first m elements of $\tilde{\boldsymbol{\beta}}$ have any effect on the penalty. Thus, the generalized sparsity penalty is operating on effectively m

variables—even if the dimensionality of $\boldsymbol{\beta}$ is p. To see this further, split $\tilde{\boldsymbol{\beta}}$ into two groups; group 1, $\tilde{\boldsymbol{\beta}}_1$, has the first m elements and group 2, $\tilde{\boldsymbol{\beta}}_2$, has the remaining $p-m$ elements. We can write the joint prior as follows.

$$p(\tilde{\boldsymbol{\beta}}_1, \tilde{\boldsymbol{\beta}}_2) = c(\lambda, \boldsymbol{D}) \exp(-\lambda ||\tilde{\boldsymbol{U}} \boldsymbol{\sigma} \tilde{\boldsymbol{\beta}}_1||) \tag{16}$$

Thus, we can note that the prior can be written as follows: There is a generalized LASSO prior on $\tilde{\boldsymbol{\beta}}_1$ but a flat prior on $\tilde{\boldsymbol{\beta}}_2$. Thus, in the $\tilde{\boldsymbol{\beta}}$ space, we have a prior that is, in fact, improper on $p-m$ elements of $\tilde{\boldsymbol{\beta}}$. For the proper prior, however, we can find its normalizing constant in terms of λ and some value $w_{\boldsymbol{D}}$ that only depends on $\boldsymbol{D}$. That is,

$$p(\tilde{\boldsymbol{\beta}}_1) = \lambda^m w_{\boldsymbol{D}} \exp(-\lambda ||\tilde{\boldsymbol{U}} \boldsymbol{\sigma} \tilde{\boldsymbol{\beta}}_1||) \tag{17}$$

From this, we can reconstruct the joint prior again, noting that proportional to reflects the fact that there is a flat prior for $p-m$ elements. We can change back into the $\boldsymbol{\beta}$ space to get the result in the main text.

$$p(\boldsymbol{\beta}) \propto \lambda^m w_{\boldsymbol{D}} \exp(-\lambda ||\boldsymbol{D} \boldsymbol{\beta}||) \tag{18}$$

Some remarks: As noted in the main text, I write the conditional prior as below where $^+$ denotes the pseudo-inverse.

$$\boldsymbol{\beta} | \{\tau_k^2\} \sim N\left(\mathbf{0}, \left(\boldsymbol{D}^T \boldsymbol{\tau}^{-1} \boldsymbol{D}\right)^+\right) \tag{19}$$

Given the results above, we can put the point more sharply. If we work in the space of $\tilde{\boldsymbol{\beta}}$, then we have a conditional prior for $\tilde{\boldsymbol{\beta}}_1$ that is always well-defined with a positive-definite variance:

$$\tilde{\boldsymbol{\beta}}_1 | \{\tau_k^2\} \sim N\left(\mathbf{0}, (\tilde{\boldsymbol{D}}^T \boldsymbol{\tau} \tilde{\boldsymbol{D}})^{-1}\right) \tag{20}$$

Second, to get the representation in terms of the pseudo-inverse, we can define $\check{\boldsymbol{\beta}} \sim N(0, \boldsymbol{S}^{-1})$ where $\boldsymbol{S}^{-1}$ is the diagonal matrix formed by the reciprocal of the non-zero eigenvalues of the

eigen-decomposition of $\boldsymbol{D}^T\boldsymbol{\tau}^{-1}\boldsymbol{D}$. We can thus cast the prior as follows:

$$\check{\boldsymbol{\beta}}|\{\tau_k^2\} \sim N(\mathbf{0}_{m\times 1}, \boldsymbol{S}^{-1}) \tag{21a}$$

$$\boldsymbol{\beta} = \boldsymbol{Q}^T \begin{pmatrix} \boldsymbol{I} \\ \mathbf{0} \end{pmatrix} \check{\boldsymbol{\beta}} \tag{21b}$$

Thus, if we take a full rank $\check{\boldsymbol{\beta}}$ generated from the appropriate multivariate normal characterized by the eigen-values of the lower dimensional space, and then project it into a larger k-dimensional space, we recover a density that can be expressed as in the main text: To note, this density can be written as follows, noting the restrictions on the support of $\boldsymbol{\beta}$

$$p(\boldsymbol{\beta}) = det^*(2\pi\boldsymbol{D}^T\boldsymbol{\tau}\boldsymbol{D})^{1/2}\exp\left(-1/2\boldsymbol{\beta}^T\boldsymbol{D}\boldsymbol{\tau}^{-1}\boldsymbol{D}\right) \tag{22}$$

This has important implications for the joint density: Recall that I proposed the following joint prior for the Bayesian generalized LASSO:

$$p(\boldsymbol{\beta}, \{\tau_k^2\}) \propto \lambda^{m+k} det(2\pi\boldsymbol{\tau})^{-1/2}\exp(-\lambda^2/2\sum\tau_k^2)\exp\left(-1/2\boldsymbol{\beta}^T\boldsymbol{D}\boldsymbol{\tau}^{-1}\boldsymbol{D}\right) \tag{23}$$

Integrating away $\boldsymbol{\beta}$, over the space on which it is defined, gives the following *marginal* density for $\{\tau_k^2\}$:

$$p(\{\tau_k^2\}) \propto det^*(2\pi\boldsymbol{D}^T\boldsymbol{\tau}^{-1}\boldsymbol{D})^{-1/2} det^*(2\pi\boldsymbol{\tau})^{-1/2}\exp(-\lambda^2/2\sum_k\tau_k^2)\lambda^{m+k} \tag{24}$$

Some remarks are in order. First, this density can in only special circumstances be written as independent exponential distributions as in the vanilla case. Specifically, only when $rank(\boldsymbol{D}) = K$. In this case, we note that $det^*(\boldsymbol{D}^T\boldsymbol{\tau}^{-1}\boldsymbol{D}) = det^*(\boldsymbol{\tau}^{-1}\boldsymbol{D}\boldsymbol{D}^T)$. Since both τ^{-1} and $\boldsymbol{D}\boldsymbol{D}^T$ are invertible, this means that $det^*(\boldsymbol{D}^T\boldsymbol{\tau}^{-1}\boldsymbol{D}) = det(\boldsymbol{D}\boldsymbol{D}^T)\prod_k(1/\tau_k^2)^{-1}$. Thus, we can write the marginal density in this case as follows; a product of independent exponential distributions.

$$p(\{\tau_k^2\}) \propto \prod_k \lambda^2/2\exp(-\lambda^2/2\tau_k^2) \tag{25}$$

In the case of the categorical sparsity and other case where $rank(\boldsymbol{D}) = p$, we have a proper prior on all elements of $\boldsymbol{\beta}$. This can also be guaranteed by adding a vanilla sparsity to all elements of

$\boldsymbol{\beta}$ regardless of the pre-existing $\boldsymbol{D}$. By a similar logic, the prior on $\boldsymbol{\beta}$ can always be made proper by adding a (proper) normal prior to all elements in addition to the complex sparsity. This, however, has a somewhat undesirable property insofar as it pulls all elements (slightly) towards zero which may be in tension with the desired type of sparsity (e.g. temporal or functional). One could address this by adding the diffuse normal prior to $\tilde{\boldsymbol{\beta}}_2$. Working through the math gives the following (proper) prior; assume that $p(\tilde{\boldsymbol{\beta}}_2) \sim N(0, \boldsymbol{\Sigma}_{0,\beta})$.

$$p(\boldsymbol{\beta}) = \lambda^m w_{\boldsymbol{D},\boldsymbol{\Sigma}_{0,\beta}} det(2\pi\boldsymbol{\Sigma}_{0,\beta})^{-1/2} \exp\left(-\lambda||\boldsymbol{D}\boldsymbol{\beta}|| - 1/2\boldsymbol{\beta}\boldsymbol{V}\begin{pmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{0,\beta}^{-1}\end{pmatrix}\boldsymbol{V}^T\boldsymbol{\beta}\right) \tag{26}$$

As $\boldsymbol{\Sigma}_{0,\beta}$ tends to infinity, i.e. becomes infinitely diffuse, the limiting case is the one noted in the main text.

Finally, it is worth also considering the cases where the posterior on $\boldsymbol{\beta}$ is well defined. First, it is obvious to note that if either the prior is proper or $\boldsymbol{X}$ is full rank, then the posterior is proper. Considering the conditional posterior, its precision can be expressed as follows. $\left(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{D}^T\boldsymbol{\tau}^{-1}\boldsymbol{D}\right)$. Does this invert? In the case of either $\boldsymbol{X}$ or $\boldsymbol{D}$ being full column rank, it does. More generally, if $\begin{pmatrix}\boldsymbol{X} \\ \boldsymbol{D}\end{pmatrix}$ is full column rank, then it does. For the cases considered above, many types of complex sparsity involve constructing $\boldsymbol{X}$ such that the relevant (penalized) columns have full rank, e.g. creating a 'one hot' structure. More generally, Ali and Tibshirani (2018) outline cases in which the (frequentist) generalized LASSO is unique; future work will examine whether it can be proved in those cases that for an arbitrary $\boldsymbol{X}$, the posterior is proper. It is also worth noting that adding a diffuse ridge prior to non-penalized elements in addition to the diffuse prior on the penalized elements (e.g. via $\boldsymbol{\Sigma}_{0,\beta}$) can always guarantee a proper posterior as it gives a fully proper prior to all elements of $\boldsymbol{\beta}$—penalized or not.

B Details of the Other Forms of LASSO

B.1 Specific Generalized LASSOs

As noted in the main text, the functional, geographic, and categorical LASSO are particular instantiations of the generalized LASSO. The specific forms are described below:

- Vanilla LASSO: In this case, $\boldsymbol{D} = \boldsymbol{I}$.

- Fused LASSO: As outlined above, this consists of a matrix that has the following shape.

$$\boldsymbol{D} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix}$$

  More formally, for each $t \leq T-1$, there should be one row of $\boldsymbol{D}$ that has a '1' in the column position corresponding to $\beta_{j,t}$ and a '-1' to the $\beta_{j,t+1}$. Permuting the rows to have the structure above, i.e. ordering $\boldsymbol{\beta}_j$ from $t = 1$ to T, gives $\sum_k \boldsymbol{d}_k \boldsymbol{d}_k^T / \tau_k^2$ a tridiagonal structure.

- Functional: The exact formula for a penalty of some polynomial order k is as follows (Tibshirani 2014). For some $\beta_{j,t}$ where $t \leq T-k-1$, create a row of $\boldsymbol{D}$ where the columns corresponding to the elements $\beta_{j,s} \quad s \in \{t, \cdots, t+k+1\}$ have the following values:

$$(-1)^{s-t} \binom{k+1}{s-t}$$

  Put another way, for some order k, one takes the relevant coefficients from a binomial expansion and places them starting from position t to $t+k+1$ with alternating signs. The temporal LASSO can be thought of as a '0-order' polynomial on the values—i.e. no penalty if there is *no* change—as it produces a piecewise constant function. For clarity, the first two rows for $k \in \{1, 2, 3\}$ are shown below assuming $\boldsymbol{\beta}_j$ is ordered from $t = 1$ to T. Thus, note that if the time series is length T, the largest polynomial that can be estimated is $T-k-1$.

  1. $k = 1$:
  $$\boldsymbol{D} = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \end{pmatrix}$$
  2. $k = 2$:
  $$\boldsymbol{D} = \begin{pmatrix} 1 & -3 & 3 & -1 & 0 \\ 0 & 1 & -3 & 3 & -1 \end{pmatrix}$$
  3. $k = 3$:
  $$\boldsymbol{D} = \begin{pmatrix} 1 & -4 & 6 & -4 & 1 & 0 \\ 0 & 1 & -4 & 6 & -4 & 1 \end{pmatrix}$$

- Geographic: For each geographic unit s, assume that its neighbors are denoted by ∂_s. We create one restriction for each neighbor such that the column corresponding to $\beta_{j,s}$ is given a '1' and the column for each neighbor $s' \in \partial_s$ is given a '1'. It is thus the temporal LASSO penalty for all neighboring nodes. It is almost certain in this case that $\boldsymbol{D}$ has more rows than columns and, also, is not full column rank. However, given the limited nature of the geographic dependencies, it is almost surely possible to re-arrange $\sum_k \boldsymbol{d}_k \boldsymbol{d}_k^T / \tau_k^2$ to be a banded matrix with a much smaller band than the number of columns of $\boldsymbol{D}$. As Rue (2001) and Bhattacharya, Chakraborty, and Mallick (2016) show, this allows for fast inference.

- Categorical: We have some indicator variable $\boldsymbol{\beta}_j$ with C levels. We create a zero-one indicator for all levels C, i.e. *with no baseline*. We thus structure $\boldsymbol{D}$ to impose a temporal LASSO ('1', '-1') penalty between all unique pairs of levels. This leads to $C(C-1)/2$ restrictions.

As noted above, any of these can be mixed with group or vanilla penalties. For example, one can create a categorical sparse LASSO by mixing a vanilla LASSO penalty with the categorical restrictions. This would encourage all levels to both equal to each other and equal to zero. The levels that 'remained' therefore would be clustered (i.e. sharing a small number of common values) as well as mostly being sparse (i.e. most levels being zero). In general, adding a vanilla penalty involves appending an identity matrix $\boldsymbol{I}$ with the dimensionality of $\boldsymbol{\beta}_j$ to the bottom of $\boldsymbol{D}$.

Adding in the group LASSO requires more conceptual work insofar as it is easy to write the penalty on the augmented scale but it cannot be simply added to $\boldsymbol{D}$ directly. Assume we wanted to add a group penalty (Yuan and Lin 2006) to some generalized LASSO penalty. Following Kyung et al. (2010), it has a Bayesian representation, where p_j represents the dimensionality of $\boldsymbol{\beta}_j$:

$$p(\boldsymbol{\beta}_j) \propto \exp\left(-\lambda_1 ||\boldsymbol{D}\boldsymbol{\beta}_j||_1 - \lambda_2 \sqrt{\boldsymbol{\beta}_j^T \boldsymbol{F} \boldsymbol{\beta}_j}\right) \tag{27a}$$

$$\boldsymbol{\beta}_j \sim N\left(0, \sum_k \tau_k^2 \boldsymbol{d}_k \boldsymbol{d}_k^T + \xi_g \boldsymbol{F}^{-1}\right); \quad \tau_k^2 \sim^{iid} \text{Exp}(\lambda_1^2/2); \xi_g \sim \Gamma([p_j+1]/2, \lambda_2^2/2) \tag{27b}$$

This nicely illustrates the fundamental difference between the group and generalized LASSO. There is no $\boldsymbol{d}_k \boldsymbol{d}_k^T$ such that we get, say, the identity matrix.[27] Further, even if we have the vanilla LASSO, we have a diagonal matrix with *independent* τ_k^2 governing each element whereas the group LASSO has a single augmentation variable (ξ_g^2) governing the *entire* matrix.

[27] For a simple proof, $\boldsymbol{d}_k \boldsymbol{d}_k^T$ gives us a matrix where each ij is $[\boldsymbol{d}_k]_i [\boldsymbol{d}_k]_j$. The identity matrix requires that all diagonal elements be '1' and all off diagonal elements be '0'. Outside of the degenerate case where each group is size one (and we recover the vanilla LASSO), we see that this is not possible.

Further extensions on the group LASSO, e.g. the temporal group LASSO, are also possible to mix with the generalized LASSO by noting that we simply need to combine the kernels of two multivariate normals. As that is not the focus of the paper, I do not derive that proof here, but it is roughly analogous to the derivation of the temporal sparse LASSO.

B.2 Banded Generalized LASSO

Inference for a generalized LASSO with arbitrary $\boldsymbol{D}$ can be conducted in two ways. First, one can simply invert the (weighted) outer product of the data plus the precision matrix each time. To speed this up, I note a number of important 'tricks'.

First, one should think of $\boldsymbol{D}$ as being specific to *each variable* and thus we can optimize them coordinate-wise. That is, if we are simply doing a vanilla LASSO, it is almost surely faster to cycle through the coordinates one-at-a-time versus doing a full inversion especially in the case where $P >> N$. This thus implies that if we are doing, say, a temporal LASSO for each variable, we need to do a temporal LASSO update for each coordinate versus stacking together the entire matrix.

To illustrate this, consider an example from text-as-data where we tried to predict party based on 6,000 words across 50 Congresses. That would be a very large (300,000) matrix to invert to do all of the updates directly and even the speedup techniques by Bhattacharya, Chakraborty, and Mallick (2016) and Johndrow, Orenstein, and Bhattacharya (2018) were employed, would still likely run into severe difficulty. By contrast, if we focus on doing each word sequentially, we have merely a 50 by 50 matrix to invert that we need to do 6,000 times. The downside of the cyclical approach, however, is that the residualization required for the update itself requires some matrix multiplications; however, if that proves too costly, other approximate strategies may be employed.

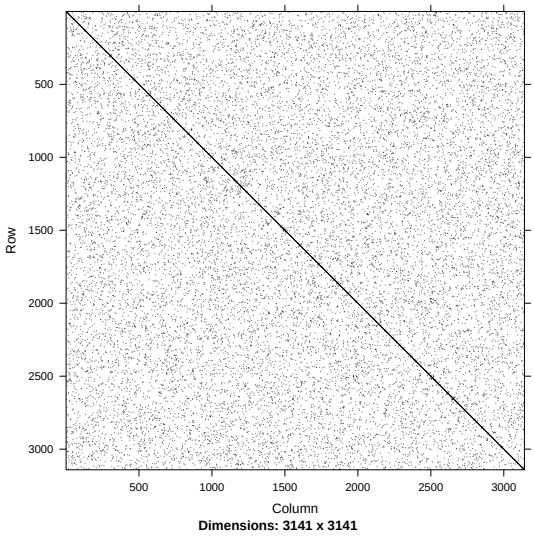
Second, it is likely possible to do even better for a specific $\boldsymbol{D}_j$ than "naive" inversion insofar as they are likely *banded*. The banded structure is automatic in the case of functional and temporal sparsity and appears in many cases of geographic sparsity. As an example, consider the map of the counties in the United States. I extract the corresponding to neighbors for each county—with each county considered to be its own neighbor and plot the adjacency matrix using the ordering of the counties provided in the shapefile. We see that this matrix looks very complicated and not amenable to 'simple' tricks for fast inversion.

However, following Rue (2001), there is a large literature on how to permute the rows and columns of a sparse matrix to create a banded matrix with the smallest (or much smaller) bandwidth.

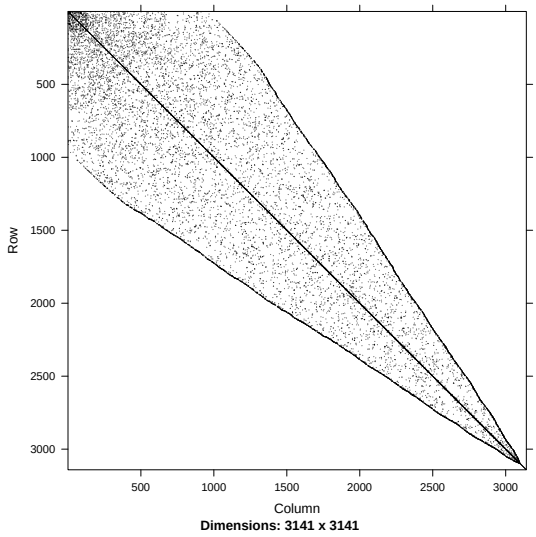
Using a variety of algorithms, one can re-order the counties to get the following adjacency matrix. This is *identical* to the one shown previously—I simply reordered the counties while keeping all links. We see that this has a bandwidth of 3114 (the number of counties) versus 958 below.

Figure 4: US County Adjacency Matrix

(a) Untransformed Graph



(b) Transformed Graph



The smaller bandwidth is highly important because there are specialized routines for computing the Cholesky decomposition of a sparse banded matrix that scale on the order of mb^2 where b is the bandwidth and m is the number of nodes. This is extremely useful for evaluating quantities that require the inverse such as the log-posterior. Thus, I would always suggest for a large $\boldsymbol{D}$ and the corresponding sparse data matrix, permuting the matrix to have a small bandwidth—permuting the columns of the data to match—and then using that in one's analysis.

C Details of Inference

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):

$$\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2 \sim N\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_N\right) \tag{28a}$$

$$p(\boldsymbol{\beta}|\lambda^2, \sigma^2) \propto w_{\boldsymbol{D}} \lambda^m / \sigma^m \exp\left(-\frac{\lambda}{\sigma}||\boldsymbol{D}\boldsymbol{\beta}||_1\right) \tag{28b}$$

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 only involving $\boldsymbol{D}$:

$$-\frac{N}{2}\ln(2\pi\sigma^2) - \frac{||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2}{2\sigma^2} + m\ln(\lambda) - m/2\ln(\sigma^2) - \frac{\lambda}{\sigma}||\boldsymbol{D}\boldsymbol{\beta}||_1 + \ln p_0(\sigma^2) + \ln p_0(\lambda^2) \tag{29}$$

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, with $p(\boldsymbol{\tau})$ following the same marginal as in Appendix A.

$$p(\boldsymbol{\beta}, \boldsymbol{\tau}|\lambda, \sigma^2) \propto \sigma^{-m}\lambda^{K+m} det(\boldsymbol{\tau})^{-1/2} \exp(\sum_k -\lambda^2/2\tau_k^2) \exp\left(-\frac{\boldsymbol{\beta}^T \boldsymbol{D}^T \boldsymbol{\tau}^{-1} \boldsymbol{D}\boldsymbol{\beta}}{2\sigma^2}\right) \tag{30}$$

In the case where $\boldsymbol{D}$ is rank K, this simplifies to the familiar form:

$$\boldsymbol{\beta}|\boldsymbol{\tau}, \boldsymbol{\sigma}^2 \sim N(0, \sigma^2[\boldsymbol{D}^T\boldsymbol{\tau}\boldsymbol{D}]^+); \tau_k^2 \sim Exp(\lambda^2/2) \tag{31}$$

Recalling the results above, I note that $\boldsymbol{D}^T\boldsymbol{\tau}^{-1}\boldsymbol{D}$ may not have full rank and thus could be re-written in terms of the rotated subspace. The rotation of the $\boldsymbol{\beta}$ and $\boldsymbol{X}$ does not affect the error variance σ^2. The full augmented log-posterior can be written therefore as

$$-\frac{N}{2}\ln(2\pi\sigma^2) - \frac{||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2}{2\sigma^2} + (K+m)/2\ln(\lambda^2) - \sum_k \tau_k^2\lambda^2/2 - 1/2\sum_k \ln(\tau_k^2) \tag{32a}$$

$$-m/2\ln(\sigma^2) - \frac{\boldsymbol{\beta}^T\boldsymbol{D}^T\boldsymbol{\tau}^{-1}\boldsymbol{D}\boldsymbol{\beta}}{2\sigma^2} \tag{32b}$$

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

$$\begin{aligned}\sigma^2|- \sim \text{InverseGamma}\Bigg(& a_{0,\sigma} + \frac{1}{2}\left[N + Rank(\boldsymbol{D})\right], \\ & b_{0,\sigma} + \frac{1}{2}\left[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{\beta}^T\boldsymbol{D}^T\boldsymbol{\tau}^{-1}\boldsymbol{D}\boldsymbol{\beta}\right]\Bigg)\end{aligned} \tag{33}$$

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.

$$\boldsymbol{\beta}|. \sim N\left(\boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{y}, \sigma^2\boldsymbol{A}^{-1}\right); \quad \boldsymbol{A} = (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\tau}^{-1}) \tag{34a}$$

$$1/\tau_k^2 \sim \text{InvGaussian}\left(\frac{\lambda\sigma}{|\boldsymbol{d}_k^T\boldsymbol{\beta}_j|}, \lambda^2\right) \tag{34b}$$

$$\lambda^2 \sim \text{Gamma}\left(a_{0,\Lambda} + [K+m]/2, \quad b_{0,\Lambda} + \frac{1}{2}\sum_{k=1}^{K}\tau_k^2\right) \tag{34c}$$

C.2 Multinomial Regression

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

$$p(y_i = k|\{\boldsymbol{\beta}_k\}) \propto \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}_k) \tag{35}$$

We can thus write the likelihood as, setting $\boldsymbol{\beta}_K = \boldsymbol{0}$ to identify the model.

$$\prod_{i=1}^{N}\left[\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}_k)}{\sum_{l=1}^{K}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}_k)}\right]^{I(y_i=k)} \tag{36}$$

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

Assume, therefore, that the prior structure has the form

$$p(\{\boldsymbol{\beta}_k\}) \propto \prod_{k=1}^{K-1}\lambda^m\exp(-\lambda||\boldsymbol{D}\boldsymbol{\beta}_k||) \tag{37}$$

One could write this using an augmented $\boldsymbol{D}$ and stacked $\boldsymbol{\beta}$ if one wished to impose restrictions

across levels of $\boldsymbol{\beta}_k$. We can estimate this using an AECM algorithm (Meng and Van Dyk 1997)—or a Gibbs Sampler; a detailed exposition can be found in the Appendix of Goplerud (2018a). The intuition behind the AECM algorithm is that we can augment with different Polya-Gammas to cyclically update the $\boldsymbol{\beta}_k$. At each step, we perform an E-Step update for the relevant Polya-Gammas and the necessary τ^2, noting that there are now different τ^2 for each level.

Specifically, for each k, we want to do inference on the following posterior:

$$p(\{\boldsymbol{\beta}_k\}|\{\boldsymbol{\beta}_{\neg k}\}) = \prod_{i=1}^{N} \frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}_k - C_{ik})^{I(y_i=k)}}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}_k - O_{ik}) + 1} \cdot p(\{\boldsymbol{\beta}_k\}_{t=1}^T) \tag{38}$$

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

Thus, we can do Polya-Gamma augmentation as outlined in Goplerud (2018a), see also Polson, Scott, and Windle (2013). To derive the relevant E-step, note that:

$$\omega_{ik}^* = E[\omega_{ik}|\boldsymbol{\beta}_k, \boldsymbol{\beta}_{\neg k}] = \frac{1}{2\psi_{ik}}\tanh(\psi_{ik}/2); \quad \psi_{ik} = \boldsymbol{x}_i^T\boldsymbol{\beta}_k - O_{ik} \tag{39}$$

Thus, the complete data log-posterior for $\boldsymbol{\beta}_k$ conditional on $\boldsymbol{\beta}_{\neg k}$ can be written as:

$$\sum_{i=1}^{N}[I(y_i = k) - 1/2](\boldsymbol{x}_i^T\boldsymbol{\beta}_k - O_{ik}) - \omega_{ik}^*/2(\boldsymbol{x}_i^T\boldsymbol{\beta}_k - O_{ik})^2 - \lambda||\boldsymbol{D}\boldsymbol{\beta}_k \tag{40}$$

Noting that the data augmentation for the complex sparsity penalty occurs independently of the augmentation for the Polya-Gammas (ω_{ik}), this can be done simultaneously giving the following complete data log-posterior:

$$\sum_{i=1}^{N}[I(y_i = k) - 1/2](\boldsymbol{x}_i^T\boldsymbol{\beta}_k - O_{ik}) - \omega_{ik}^*/2(\boldsymbol{x}_i^T\boldsymbol{\beta}_k - O_{ik})^2 - 1/2\boldsymbol{\beta}^T\boldsymbol{D}^T[\boldsymbol{\tau}^*]\boldsymbol{D}\boldsymbol{\beta} \tag{41}$$

$$[\boldsymbol{\tau}^*]_i = E\left[\frac{1}{\tau_k^2}\right] = \frac{\lambda}{|\boldsymbol{x}_i^T\boldsymbol{\beta}_k|}$$

$\boldsymbol{\beta}_k$ can be maximized using least squares and thus the M-Step is tractable. By iterating this process for each k, we can deterministically find the posterior mode.

D Strategies for Estimating λ

This section fleshes out some practical details of using the methods outlined above.

D.1 Cross-Validation and Information Criteria

This method is conceptually straightforward in terms of what needs to be fit, but some non-trivial practical issues are worth noting. First, assume one has created a grid of λ from very small (implying little regularization) to very large (implying a large amount of regularization). We then estimate the model for each λ. If one is using cross-validation, it is simply a matter of calculating the model performance on the test-set using the coefficients from the model.

To calculate the information criteria, however, one needs to estimate the degree of freedom following Tibshirani and Taylor (2011). Exactly which restrictions are "active", i.e. $\boldsymbol{d}_k^T\boldsymbol{\beta} = 0$, depends on the numerical precision of one's computer as the EM algorithm. I thus recommend checking a number of difference tolerances with 10^{-7} being a good starting point. The associated software package rescales all non-binary $\boldsymbol{X}$ (and linear outcomes) to be on a standard deviation scale and thus 10^{-7} is "effectively" zero.

Further, in terms of "refitting" the model, again one must decide which $\boldsymbol{\beta}$ are sufficiently close to be grouped together; analogously, in the refitting case on the vanilla LASSO, which coefficients are sufficiently close to zero to be ignored. Here, I use a slightly weaker standard (10^{-4}), again noting that since the data are standardized, this is actually a very small difference in standardized terms. Again, one could examine the robustness of the results to this threshold. I conjecture that this will mostly matter in terms of collapsing very small or highly similar groups together.

In terms of actually fitting the models along the regularization path, it is necessary to go "backwards" when using the EM strategy. Noting that when $\boldsymbol{d}_k\boldsymbol{\beta} = 0$, this is a fixed point in the algorithm and that restriction *must* hold for all subsequent iterations. Thus, initializing a model at $\boldsymbol{\beta} = 0$, as in the typical vanilla case, will not work. In this case, however, the best option is to start with a small λ (little regularization) and fit the model from scratch. Then, one can increase λ slightly using the old values as starting values, and then run to convergence. By iterating this process and increasing λ from small to large, the entire path can be computed. The ideas in Park and Hastie 2007 about calculating paths for non-linear models with the vanilla LASSO can likely be adapted here for future increases in speed.

One estimates the coefficients for that fixed λ and then chooses an optimal λ based on some performance on held-out data. There are three major issues with this framework: First, naive cross-validation alone tends to select too small of λ (overfitting) and thus a variety of *ad hoc* strategies (i.e. choose a λ that is one standard error larger than the cross-validated λ) exists. Second, for multiple tuning parameters, i.e. if we have different λ governing the vanilla and

temporal sparsity, cross-validation becomes prohibitively expensive. Third, a variety of research (e.g. Park and Casella 2008; Polson and Scott 2011a) has shown that having λ be a random variable improves performance as the model can average over regularization strengths.

D.2 Gibbs Sampler

The other major approach to selecting λ is to give it a prior and sample from its full conditional. The distribution can be backed out from the posterior following Kyung et al. (2010) and the results noted above.

$$p(\lambda^2|-) \propto (\lambda^2)^{(k+rank(\boldsymbol{D}))/2} \exp(-1/2 \sum_j \tau_k^2 \lambda^2) \pi(\lambda^2) \tag{42}$$

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. As noted above, there is existing research to suggest that in the vanilla case, giving λ^2 a mean that grows at rate $N \ln P$ is reasonable. Thus, the adaptation to this case is $N \ln rank(\boldsymbol{D})$ given that Appendix A showed that we can think of there being $m = rank(\boldsymbol{D})$ penalized coefficients.

D.3 EM Algorithm

As suggested above in passing, a hybrid strategy exists that estimates λ alongside $\boldsymbol{\beta}$ simultaneously—in a deterministic fashion unlike the fully Bayesian paradigm. Assume that we have adopted the strategy above and placed a prior on λ^2. We have a posterior over $p(\boldsymbol{\beta}, \lambda^2|\boldsymbol{X}, \boldsymbol{y})$ that we wish to sample from. Typically, we have assumed λ as fixed and thus maximized $p(\boldsymbol{\beta}|\lambda, \boldsymbol{y}, \boldsymbol{X})$ with respect to $\boldsymbol{\beta}$. However, we could also maximize the joint posterior of $p(\boldsymbol{\beta}, \lambda^2|\boldsymbol{y}, \boldsymbol{X})$ with respect to *both* $\boldsymbol{\beta}$ and λ. In EM jargon, we put λ into the M-Step of the algorithm.

This is not quite the quantity of interest that we typically care about, either from the Bayesian or the frequentist paradigms, but it is one that seems to perform well in practice, see Ratkovic and Tingley (2017) and Goplerud et al. (2018), and has also been used in other cases Polson and Scott (2011b). This approach has some downsides; first, it tends to overshrink somewhat but also may sometimes degenerate, i.e. λ is estimated to be zero or 'infinity', and thus there is either no shrinkage or full shrinkage. From experiments and existing work that uses this approach, however, this seems to occur less often than one might worry about, although this approach does tend to over-shrink versus the fully Bayesian method.

Thus, if it is infeasible to use the other approaches or one simply wants the fastest way of getting a reasonable point estimate for prediction, this approach seems sensible. In terms of initializing a Gibbs Sampler, this approach is ideal as it makes sure that we start in a region of high joint-posterior density.

Another possible extension is to put λ into the E-Step of the algorithm, i.e. have a joint layer of (τ_k^2, λ) to take expectations over. This requires working with the joint posterior of $p(\tau_k^2, \lambda|\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{y})$, but in the adaptive and regular LASSO penalty cases, this is likely "doable" as it requires only a one-dimensional integral to evaluate $p(\lambda|\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{y})$. In this case, we are targeting the correct mode $p(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{y})$.

E Other Regularizations

In this section, I outline the two other forms of regularization that I integrate into my framework. This is not exhaustive, but both permit a fast EM algorithm that is only a slight modification of the results discussed above.

E.1 Adaptive LASSO

The adaptive LASSO has the following prior, following Zou (2006). Assume that $w_j = 1/|\hat{\beta}_j|$ where $\hat{\beta}_j$ is some $\sqrt{N}$-consistent estimator:

$$p(\beta_j) = \frac{\lambda w_j^\gamma}{2} \exp\left(-\lambda w_j^\gamma |\beta_j|\right) \tag{43}$$

After augmentation, we get the following:

$$\beta_j \sim N(0, \tau_j^2) \tag{44a}$$

$$\tau_j^2 \sim \text{Exp}\left(\frac{\lambda^2 [w_j^\gamma]^2}{2}\right) \tag{44b}$$

$$\lambda^2 \sim \text{Gamma}(a_0, b_0) \tag{44c}$$

$$\gamma \sim \text{GenGamma}(1, 2, 1); \quad p(\gamma) = \gamma \exp(-\gamma) \tag{44d}$$

Define $\boldsymbol{W}$ as w_j^γ stacked diagonally. We can extend the idea to the generalized LASSO, up to

a constant that does not depend on γ or λ:

$$p(\boldsymbol{\beta}) \propto \lambda^{rank(\boldsymbol{D})} w_{\boldsymbol{WD}} \exp\left(-\lambda ||\boldsymbol{WD}\boldsymbol{\beta}||\right); \quad \lambda ||\boldsymbol{WD}\boldsymbol{\beta}|| = \sum_k \lambda w_k |\boldsymbol{d}_k^T \boldsymbol{\beta}| \tag{45}$$

The complication arises because the normalizing constant for the matrix, i.e. $w_{\boldsymbol{WD}}$, depends on γ in a complicated way. We can use data augmentation, grouping λ and w_k together:

$$p(\boldsymbol{\beta}, \{\tau_k^2\}) \propto det(2\pi\boldsymbol{\tau})^{-1/2} \exp(-1/2 \boldsymbol{\beta}^T \boldsymbol{D}^T \boldsymbol{\tau}^{-1} \boldsymbol{D}\boldsymbol{\beta}) \cdot \prod_k \exp(-\lambda^2 [w_j^\gamma]^2 / 2\tau_k^2) \lambda^2 [w_j^\gamma]^2 / 2 \tag{46}$$

From here, we see both the symmetry and the difference. First, it is clear that the conditional distribution for $\boldsymbol{\beta}$ is identical to before and the full conditional on τ_k^2 is

$$1/\tau_k^2 \sim \text{InvGaus}\left(\frac{\lambda w_k^\gamma}{|\boldsymbol{d}_k^T \boldsymbol{\beta}|}, \quad \lambda^2 [w_k^\gamma]^2\right) \tag{47}$$

We see that if we were to integrate away the $\boldsymbol{\beta}$,

$$p(\{\tau_k^2\}) \propto det(2\pi\boldsymbol{\tau})^{-1/2} det(2\pi \boldsymbol{D}^T \boldsymbol{\tau}^{-1} \boldsymbol{D})^{-1/2} \cdot \prod_k \exp(-\lambda^2 [w_j^\gamma]^2 / 2\tau_k^2) \lambda^2 [w_j^\gamma]^2 / 2 \tag{48}$$

If $\boldsymbol{D}\boldsymbol{D}^T$ is invertible, i.e. $rank(\boldsymbol{D}) = K$, then we get the independent exponential distributions as the marginal on τ_k^2: We can use this result to note that

$$p(\{\tau_k^2\}) = \prod_k \exp(-\lambda^2 [w_j]^\gamma / 2\tau_k^2) \lambda^2 [w_j^\gamma]^2 / 2 \tag{49a}$$

$$p(\boldsymbol{\beta}) = (\boldsymbol{D}\boldsymbol{D}^T)^{-1} \lambda^K \cdot 2^{-K} \cdot \left[\prod_k w_k\right]^\gamma \cdot \exp(-\lambda ||\boldsymbol{WD}\boldsymbol{\beta}||) \tag{49b}$$

In this particular case, therefore, we can write the normalizing constant separately in terms of functions of $\boldsymbol{D}$, γ, and λ. However, a similar simplification is not possible for the general case.

The difference is that the adaptive LASSO left multiplies $\boldsymbol{D}$ by a diagonal matrix of the weights $w_k = |\boldsymbol{d}_k^T \hat{\boldsymbol{\beta}}|$ raised to γ power. Thus, for now, I would assume a fixed $\gamma = 1$ when using the adaptive LASSO or, perhaps, examine how varying this parameter affects the performance of the model. Further research may seek to explore whether the normalizing in terms of γ can be drawn out and, if so, whether it is possible to sample from γ directly.

Fortunately, inference on λ is unchanged as its role in the normalizing constant is unchanged. The full conditional is below, see Appendix A for details:

$$\lambda^2 \sim Gamma\left(a_0 + [K+m]/2, \quad b_0 + \frac{1}{2}\sum_j \tau_j^2 [w_j^\gamma]^2\right) \tag{50a}$$

To estimate this via the EM algorithm, if λ^2 is in the M-Step, this is effectively the same algorithm for the vanilla LASSO. The relevant moments for the EM algorithm are also listed below.

$$E[1/\tau_j^2] = \frac{\lambda w_j^\gamma}{|\beta_j|} \tag{51a}$$

$$[w_j^\gamma]^2 E[\tau_j^2] = [w_j^\gamma]^2 \left[\frac{|\beta_j|}{\lambda w_j^\gamma} + \frac{1}{\lambda^2 [w_j^\gamma]^2}\right] = \frac{|\beta_j|}{\lambda} w_j^\gamma + \frac{1}{\lambda^2} \tag{51b}$$

$$\lambda^2 [w_j^\gamma]^2 [E\tau_j^2] = \lambda |\beta_j| w_j^\gamma + 1 \tag{51c}$$

F Details on Trounstine (2016)

In the first set of models (city segregation on the racial gap in voting), the following controls are used: Diversity, % Black, % Asian, % Latino, median income, % renters, % college degree, biracial contest, nonpartisan election, primary election, and logged population.

In the second set of models (city segregation on public goods outcomes), the following controls are used: Diversity, % Black, % Asian, % Latino, median income, % local government employees, % renters, % over 65, % college degree, population logged.