Modelling Heterogeneity Using Bayesian Structured Sparsity: Supporting Information

A Proof of Theorems on Structured Sparsity

I note a number of standard results from linear algebra that are referenced in the proofs.

R1 For any $\boldsymbol{x} \in \mathbb{R}^p$, $||\boldsymbol{x}||_2 \leq ||\boldsymbol{x}||_1 \leq \sqrt{p}||\boldsymbol{x}||_2 \leq \sqrt{p}||\boldsymbol{x}||_1$

R2 Any matrix $\boldsymbol{D} \in \mathbb{R}^{n \times p}$ admits a singular value decomposition (SVD) of the following form where $\boldsymbol{\Sigma}$ is a diagonal matrix with positive entries of size $\text{rank}(\boldsymbol{D}) = m$ and $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices of size $n \times n$ and $p \times p$, respectively.

$$\boldsymbol{D} = \boldsymbol{U} \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{0}_{m \times p-m} \\ \boldsymbol{0}_{n-m \times m} & \boldsymbol{0}_{n-m \times p-m} \end{pmatrix} \boldsymbol{V}^T \tag{1}$$

A "thin" SVD is defined as $\boldsymbol{D} = \boldsymbol{U}_1 \boldsymbol{\Sigma} \boldsymbol{V}_1^T$ where the subscript '1' denotes taking the first m columns of a matrix.

R3 Assume that $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ and $\boldsymbol{D} \in \mathbb{R}^{m \times p}$. Denote $\mathcal{N}(\boldsymbol{A})$ as the nullspace of $\boldsymbol{A}$ and $\mathcal{B}_{\boldsymbol{A}}$ as a basis for this nullspace. Define the set $\mathcal{S} \subseteq \mathbb{R}^p$ such that $\mathcal{S} = \{s : \boldsymbol{X}\boldsymbol{s} = \boldsymbol{0} \text{ and } \boldsymbol{D}\boldsymbol{s} = \boldsymbol{0}\}$. The following conditions are all equivalent:

(a) $\mathcal{S} = \{\boldsymbol{0}\}$, i.e. the only member of $\mathcal{S}$ is the zero vector $\boldsymbol{0}$.

(b) $\mathcal{N}(\boldsymbol{X}) \cap \mathcal{N}(\boldsymbol{D}) = \{\boldsymbol{0}\}$

(c) $\text{rank}\left(\begin{bmatrix}\boldsymbol{X}^T & \boldsymbol{D}^T\end{bmatrix}\right) = p$

(d) $\text{rank}(\boldsymbol{X}\mathcal{B}_{\boldsymbol{D}}) = p - \text{rank}(\boldsymbol{D})$

R4 Consider the following optimization problem where $f : \mathbb{R}^p \to \mathbb{R}$ and $\boldsymbol{A} \in \mathbb{R}^{K \times p}$.

$$\boldsymbol{x}^* = \underset{\boldsymbol{x} \in \mathbb{R}^p}{\arg\max} f(\boldsymbol{x}) \quad \text{s.t.} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0} \tag{2}$$

If $\mathcal{B}_{\boldsymbol{A}}$ is a basis for the nullspace of $\boldsymbol{A}$, i.e. consisting of $p-\text{rank}(\boldsymbol{A})$ linearly independent vectors of length p, derive the following unconstrained optimization problem:

$$\boldsymbol{y}^* = \underset{\boldsymbol{y} \in \mathbb{R}^{p-\text{rank}(\boldsymbol{A})}}{\arg\max} f(\mathcal{B}_{\boldsymbol{A}}\boldsymbol{y}) \tag{3}$$

The two problems are equivalent in that characterizing the solutions to Equation 2 fully characterizes the solutions to Equation 3, and vice versa. Lawson and Hanson (1974, ch. 20) discusses this in the case of least squares, but their discussion can be immediately generalized.

A.1 Proof of Theorem 1

As noted in Theorem 1, the structured sparse prior has the following kernel where $\bar{\boldsymbol{D}}$ stacks together $\boldsymbol{D}$ and $\boldsymbol{F}_\ell$ vertically and $\boldsymbol{\beta} \in \mathbb{R}^p$.

$$k(\boldsymbol{\beta}) = \exp\left(-\lambda\left[||\boldsymbol{D\beta}||_1 + \sum_{\ell=1}^{L}\sqrt{\boldsymbol{\beta}^T\boldsymbol{F}_\ell\boldsymbol{\beta}}\right]\right) \tag{4}$$

The prior is proper when the integral of $k(\boldsymbol{\beta})$ is finite for $\lambda > 0$. This occurs if and only if $\text{rank}(\bar{\boldsymbol{D}}) = p$. The proof proceeds as follows:

First, I transform Equation 4 using a decomposition of $\boldsymbol{F}_\ell$ such that $\boldsymbol{F}_\ell = \tilde{\boldsymbol{Q}}_\ell\tilde{\boldsymbol{Q}}_\ell^T$. An eigen-decomposition provides a natural choice (R2). Thus, $\sqrt{\boldsymbol{\beta}^T\boldsymbol{F}_\ell\boldsymbol{\beta}} = ||\tilde{\boldsymbol{Q}}_\ell^T\boldsymbol{\beta}||_2$. Using the bounds on the ℓ_2 norm by the ℓ_1 norm (R1), the prior kernel can be bounded.

$$\exp\left(-\lambda\left[||\boldsymbol{D\beta}||_1 + \sum_{\ell=1}^{L}||\tilde{\boldsymbol{Q}}_\ell^T\boldsymbol{\beta}||_1\right]\right) \leq k(\boldsymbol{\beta}) \leq \exp\left(-\lambda\left[||\boldsymbol{D\beta}||_1 + \frac{1}{\sqrt{p}}\sum_{\ell=1}^{L}||\tilde{\boldsymbol{Q}}_\ell^T\boldsymbol{\beta}||_1\right]\right) \tag{5a}$$

$$\begin{aligned}&\exp\left(-\lambda||\check{\boldsymbol{D}}\boldsymbol{\beta}||_1\right) \leq k(\boldsymbol{\beta}) \leq \exp\left(-\lambda||\boldsymbol{W}\check{\boldsymbol{D}}\boldsymbol{\beta}||_1\right)\\ &\check{\boldsymbol{D}} = \left[\boldsymbol{D}^T, \tilde{\boldsymbol{Q}}_\ell, \cdots\right]^T \quad \boldsymbol{W} = \text{bdiag}\left(\boldsymbol{I}_K, \frac{1}{\sqrt{p}}\boldsymbol{I}_{L\times p}\right)\end{aligned} \tag{5b}$$

Thus, examining the behavior of a structured sparse prior with only a penalty of $\check{\boldsymbol{D}}$ is sufficient to understand the behavior of a prior with $L > 0$ as $\boldsymbol{W}$ is always full rank.

Consider the integral of the lower bound in two cases: First, assume that $\text{rank}\left(\check{\boldsymbol{D}}\right) = p$. In this case, the integral is finite by upper-bounding (R1) and noting that $\check{\boldsymbol{D}}^T\check{\boldsymbol{D}}$ is full rank such that the integral is finite as an (invertible) change of variables could be applied by eigen-decomposing $\check{\boldsymbol{D}}^T\check{\boldsymbol{D}}$.

$$\int \exp\left(-\lambda||\check{\boldsymbol{D}}\boldsymbol{\beta}||_1\right)d\boldsymbol{\beta} \leq \int \exp\left(-\lambda||\check{\boldsymbol{D}}\boldsymbol{\beta}||_2\right)d\boldsymbol{\beta} = \int \exp\left(-\lambda\sqrt{\boldsymbol{\beta}^T\check{\boldsymbol{D}}^T\check{\boldsymbol{D}}\boldsymbol{\beta}}\right)d\boldsymbol{\beta} < \infty \tag{6}$$

Thus, $\text{rank}(\check{\boldsymbol{D}}) = p$ is sufficient for posterior propriety as $\boldsymbol{W}\check{\boldsymbol{D}}$ has the same rank as $\check{\boldsymbol{D}}$. Necessity can be proved in a similar way; assume that $\text{rank}(\check{\boldsymbol{D}}) \neq p$. In this case, the following integral is infinite.

$$\int \exp\left(-\lambda\left\|\check{\boldsymbol{D}}\boldsymbol{\beta}\right\|_1\right)d\boldsymbol{\beta} = \int \exp\left(-\lambda\left\|\check{\boldsymbol{U}}\begin{pmatrix}\check{\boldsymbol{\Sigma}} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}\boldsymbol{\theta}\right\|_1\right)d\boldsymbol{\theta} \tag{7a}$$

$$= \int_{\mathbb{R}^{p-\text{rank}(\check{\boldsymbol{D}}))}}\left[\int_{\mathbb{R}^{\text{rank}(\check{\boldsymbol{D}})}} \exp\left(-\lambda||\check{\boldsymbol{U}}_1\check{\boldsymbol{\Sigma}}\boldsymbol{\theta}_{\mathcal{C}}||_1\right)d\boldsymbol{\theta}_{\mathcal{C}}\right]d\boldsymbol{\theta}_{\mathcal{N}} = \infty \tag{7b}$$

The key move is to rotate $\boldsymbol{\beta}$ via an (invertible) transformation by multiplication such that $\boldsymbol{\theta} = \check{\boldsymbol{V}}^T\boldsymbol{\beta}$ where $\check{\boldsymbol{V}}$ is the right singular matrix coming from an SVD of $\check{\boldsymbol{D}}$ and $\check{\boldsymbol{U}}$ and

$\check{\boldsymbol{\Sigma}}$ defined as in R2. For notation, denote $\boldsymbol{\theta}_{\mathcal{C}}$ as the first rank$(\check{\boldsymbol{D}})$ elements of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_{\mathcal{N}}$ as the remaining elements. Note that since $\check{\boldsymbol{D}}$ is not full rank, the dimensionality of $\boldsymbol{\theta}_{\mathcal{N}}$ is at least one. Thus, the integral diverges with respect to $\boldsymbol{\theta}_{\mathcal{N}}$ but is finite with respect to $\boldsymbol{\theta}_{\mathcal{C}}$ from the above discussion as $\check{\boldsymbol{U}}_1\check{\boldsymbol{\Sigma}}$ is full column rank. Thus, rank $\left(\check{\boldsymbol{D}}\right) = p$ is a necessary condition.

Taken together, rank$(\check{\boldsymbol{D}}) = p$ is thus necessary and sufficient for posterior propriety. Because of the bounds above, this thus describes behavior for the case of $\bar{\boldsymbol{D}}$. In that case, the normalizing constant can be expressed as $\lambda^p c$ where c is a finite constant that depends only on $\bar{\boldsymbol{D}}$.

Theorem 1 restates this more cleanly in terms of $\boldsymbol{F}_\ell$. Note that since $\boldsymbol{F}_\ell$ and $\tilde{\boldsymbol{Q}}_\ell^T$ have the same nullspace and rank, $\bar{\boldsymbol{D}}$ and $\check{\boldsymbol{D}}$ do likewise, and thus Theorem 1 follows in the form expressed in the main text.

A.2 Proof of Theorem 2

The theorem assumes an un-normalized posterior of the following form where the likelihood is assumed to be log-concave with respect to $\boldsymbol{\eta}$.

$$k(\boldsymbol{\beta}|\boldsymbol{y}) = L(\boldsymbol{\eta}|\boldsymbol{y}) \exp\left(-\lambda\left[||\boldsymbol{D}\boldsymbol{\beta}||_1 + \sum_{\ell=1}^{L}\sqrt{\boldsymbol{\beta}^T\boldsymbol{F}_\ell\boldsymbol{\beta}}\right]\right) \quad \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} \tag{8}$$

The theorem can be established by a similar method to the proof of prior propriety; again note the posterior kernel can be upper and lower bounded as follows:

$$L(\boldsymbol{\eta}|\boldsymbol{y})\exp\left(-\lambda||\check{\boldsymbol{D}}\boldsymbol{\beta}||_1\right) \leq k(\boldsymbol{\beta}|\boldsymbol{y}) \leq L(\boldsymbol{\eta}|\boldsymbol{y})\exp\left(-\lambda||\boldsymbol{W}\check{\boldsymbol{D}}\boldsymbol{\beta}||_1\right) \tag{9}$$

Thus, it suffices to examine whether the posterior with a structured sparse prior on only $\check{\boldsymbol{D}}$ and $L = 0$ is proper. I thus consider the following posterior kernel in the subsequent analysis

$$k'(\boldsymbol{\beta}|\boldsymbol{y}) = L(\boldsymbol{\eta}|\boldsymbol{y})\exp\left(-\lambda||\check{\boldsymbol{D}}\boldsymbol{\beta}||_1\right) \tag{10}$$

To begin, I perform the above transformation to orthogonally rotate $\boldsymbol{\beta}$ by the right singular matrix from an SVD of $\check{\boldsymbol{D}}$: $\boldsymbol{\theta} = \check{\boldsymbol{V}}^T\boldsymbol{\beta}$. As noted in the main text, since this transformation is invertible, posterior inference on $\boldsymbol{\beta}$ is equivalent to performing inference on $\boldsymbol{\theta}$. Thus, the posterior on $\boldsymbol{\theta}$ can be expressed as follows where $\check{\boldsymbol{X}}_1$ and $\check{\boldsymbol{X}}_2$ represent the first rank$(\check{\boldsymbol{D}})$ and remaining $p - \text{rank}(\check{\boldsymbol{D}})$ columns of the rotated design $(\boldsymbol{X}\check{\boldsymbol{V}})$.

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto L(\boldsymbol{\nu}'|\boldsymbol{y})\exp\left(-\lambda||\check{\boldsymbol{U}}_1\check{\boldsymbol{\Sigma}}\boldsymbol{\theta}_{\mathcal{C}}||_1\right) \quad \boldsymbol{\nu}' = \check{\boldsymbol{X}}_1\boldsymbol{\theta}_{\mathcal{C}} + \check{\boldsymbol{X}}_2\boldsymbol{\theta}_{\mathcal{N}} \quad \check{\boldsymbol{X}} = \boldsymbol{X}\check{\boldsymbol{V}} \tag{11}$$

Establishing the propriety of the posterior in Equation 11 can be done using results from Michalak and Morris (2016). The paper provides a number of critical results summarized below as the following lemma:

Lemma 1 *Michalak and Morris (2016): Assume a likelihood $L(\boldsymbol{\eta}|\boldsymbol{y})$ such that $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$. Define an exponentiated norm bound (ENB) as follows: An ENB holds if constants $c_0, c_1 > 0$*

exist such that[1]

$$L(\boldsymbol{\eta}|\boldsymbol{y}) \leq c_0 \exp\left(-c_1||\boldsymbol{\eta}||\right) \tag{12}$$

The following results hold:

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

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

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

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

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

Second, note that the results are stated in terms of the existence and uniqueness of the MLE on $\boldsymbol{\eta}$ *not* $\boldsymbol{\beta}$. This is designed to deal with the case of a rank deficient $\boldsymbol{X}$; adapting an example from Michalak and Morris (2016, p. 552), imagine that $\boldsymbol{X}$ had two identical columns. The MLE of $\boldsymbol{\beta}$ is clearly not unique although the MLE of $\boldsymbol{\eta}$ could be—as any MLE leads to the same $\boldsymbol{X\beta}$.

Returning to the structured sparse case, note that Equation 11 reflects the scenario described in Remark (3) of Lemma 1 where $\boldsymbol{\theta}_\mathcal{C}$ has a proper prior and there is a flat prior on $\boldsymbol{\theta}_\mathcal{N}$. I prove the following Lemma:

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

Lemma 2 *For the likelihood described in Equation 11, define $\hat{\boldsymbol{\theta}}_{\mathcal{N}}$ as follows.*

$$\hat{\boldsymbol{\theta}}_{\mathcal{N}} = \arg\max_{\boldsymbol{\gamma}} \ln L(\boldsymbol{\psi}|\boldsymbol{y}); \quad \boldsymbol{\psi} = \check{\boldsymbol{X}}_2\boldsymbol{\gamma} \tag{13}$$

If $\hat{\boldsymbol{\theta}}_{\mathcal{N}}$ exists and is unique, then the posterior on $p(\boldsymbol{\theta}|\boldsymbol{y})$ is proper.

This can be proved by directly applying Michalak and Morris (2016)'s results summarized in Lemma 1. If the MLE on $\boldsymbol{\theta}_{\mathcal{N}}$ exists and is unique, then this ensures that the MLE on $\boldsymbol{\nu}'$ exists and is unique for $\boldsymbol{\theta}_{\mathcal{C}} = 0$ as it is for $\boldsymbol{\psi}$. Thus, Remarks (1) and (2) from Lemma 1 apply and the posterior is proper for $\boldsymbol{\theta}_{\mathcal{C}} = \boldsymbol{0}$.

Note, however, that for any choice of $\boldsymbol{\theta}_{\mathcal{C}} \in \mathbb{R}^{\mathrm{rank}(\check{\boldsymbol{D}})}$, the posterior remains proper. Since the MLE of $\boldsymbol{\nu}'$ exists and is unique when $\boldsymbol{\theta}_{\mathcal{C}} = \boldsymbol{0}$, it can be simply shifted to account for a non-zero offset.[2] Thus, Remark (3) from Lemma 1 applies as there is a proper prior on $\boldsymbol{\theta}_{\mathcal{C}}$ by Theorem 1 and noting that $\check{\boldsymbol{U}}_1\check{\boldsymbol{\Sigma}}$ is full column rank. Thus, the posterior on $p(\boldsymbol{\theta}|\boldsymbol{y})$ is proper and thus a structured sparse prior on $\boldsymbol{\beta}$ with $\check{\boldsymbol{D}}$ is proper as $k'(\boldsymbol{\beta}|\boldsymbol{y})$ has a finite integral. As this (or a finite transformation) upper bounds the original posterior kernel $k(\boldsymbol{\beta}|\boldsymbol{y})$, this also ensures the original posterior on $\boldsymbol{\beta}$ is proper.

Condition (b) in Theorem 2 expresses the claim in Lemma 2 slightly differently. It states that if $\hat{\boldsymbol{\beta}}_{\mathcal{N}(\boldsymbol{D})}$, as defined below, exists and is unique, then the posterior is proper.

$$\hat{\boldsymbol{\beta}}_{\mathcal{N}(\boldsymbol{D})} = \arg\max_{\boldsymbol{\beta}} L(\boldsymbol{\eta}|\boldsymbol{y}) \quad \text{s.t.} \quad \bar{\boldsymbol{D}}\boldsymbol{\beta} = \boldsymbol{0} \tag{14}$$

The equivalence between this condition and the one defined in Lemma 2 follows in three parts. First, note that $\hat{\boldsymbol{\theta}}_{\mathcal{N}}$ can be expressed as the optimization over the entire $\boldsymbol{\theta}$ space subject to a linear constraint that $\boldsymbol{\theta}_{\mathcal{C}} = \boldsymbol{0}$. Equation 15 restates the condition in Lemma 2.

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\mathbb{R}^p} \ln L(\boldsymbol{\nu}'|\boldsymbol{y}); \quad \boldsymbol{\nu}' = \check{\boldsymbol{X}}_1\boldsymbol{\theta}_{\mathcal{C}} + \check{\boldsymbol{X}}_2\boldsymbol{\theta}_{\mathcal{N}} \quad \boldsymbol{\theta}_{\mathcal{C}} = \boldsymbol{0} \tag{15}$$

Second, note that since $\boldsymbol{\theta} = \check{\boldsymbol{V}}^T\boldsymbol{\beta}$, the problem can be rotated back into the $\boldsymbol{\beta}$ space although the optimization is now over a linear subspace defined by the span of the columns of $\check{\boldsymbol{V}}_2$ as $\boldsymbol{\beta} = \check{\boldsymbol{V}}_1\boldsymbol{\theta}_{\mathcal{C}} + \check{\boldsymbol{V}}_2\boldsymbol{\theta}_{\mathcal{N}} = \check{\boldsymbol{V}}_2\boldsymbol{\theta}_{\mathcal{N}}$ where $\boldsymbol{\theta}_{\mathcal{N}} \in \mathbb{R}^{\mathrm{rank}(\check{\boldsymbol{D}})}$. This, however, is exactly the nullspace of $\check{\boldsymbol{D}}$. By noting the equivalence of optimization over the nullspace and a model with a linear constraint (R4), Equation 16 follows. Since $\check{\boldsymbol{D}}$ and $\bar{\boldsymbol{D}}$ have equivalent nullspaces, the phrasing in Theorem 2 follows.

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \ln L(\boldsymbol{\nu}|\boldsymbol{y}); \quad \boldsymbol{\nu} = \boldsymbol{X}\boldsymbol{\beta} \quad \text{s.t.} \quad \check{\boldsymbol{D}}\boldsymbol{\beta} = \boldsymbol{0} \tag{16}$$

Finally, condition (a)—the necessary condition—follows easily by examining the rank of $\check{\boldsymbol{X}}_2$. This must be full rank for the posterior to be proper; the proof proceeds by contradiction: Assume that $\mathrm{rank}(\check{\boldsymbol{X}}_2) < p - \mathrm{rank}(\boldsymbol{D})$, i.e. it was not full column rank. By the same logic that Theorem 1 is shown to be necessary, it is clear that the integral of the kernel in Equation 11 diverges in this case as after an orthogonal rotation of $\boldsymbol{\theta}_{\mathcal{N}}$ by the singular value

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

decomposition of $\check{\boldsymbol{X}}_2$. This is because after such a rotation, there are some elements of the rotated $\boldsymbol{\theta}_{\mathcal{N}}$ that appear nowhere in the prior nor the likelihood.

Recall that $\check{\boldsymbol{X}}_2$ equals $\boldsymbol{X}$ times the basis for a nullspace of $\check{\boldsymbol{D}}$. The condition that it is full rank is thus equivalent to ensuring that the nullspaces of $\check{\boldsymbol{D}}$ and $\boldsymbol{X}$ intersect only at $\mathbf{0}$ or that their stacked matrix is full rank (R3). This is thus the condition as stated in Theorem 2 again noting that $\check{\boldsymbol{D}}$ and $\bar{\boldsymbol{D}}$ have the same nullspace.

It can also be derived directly by rotating $\boldsymbol{\beta}$ by the right singular matrix coming from the stacked SVD of $\boldsymbol{X}$ and $\bar{\boldsymbol{D}}$. In that case, if it is not full rank, there exist some components in the rotated space that no longer appear in the posterior and thus the integral over those components diverges.

A.2.1 Proof of Corollary 1

Corollary 1 can be proven straightforwardly. For the linear model, Lawson and Hanson (1974, ch. 20) contains the basic idea. Assume Condition (a) holds. This implies that $\check{\boldsymbol{X}}_2$ is full rank. For a linear model, that ensures a single unique MLE and thus (a) implies (b). As (b) is sufficient for propriety, (a) is necessary and sufficient for posterior propriety. (a) and (b) is thus a slightly redundant way of stating this claim.

For the multinomial case with a standard link (logit or probit), results in Speckman, Lee, and Sun (2009) can be employed. Specifically, their Theorem 3 restated as a lemma notes:

Lemma 3 *Speckman, Lee, and Sun (2009, p. 742): For the multinomial logistic or probit choice model, the following conditions are equivalent.*

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

For Corollary 1, assume that Conditions (a) and (b) hold. This implies that $\check{\boldsymbol{X}}_2$ is full rank. Thus, if the MLE of $\hat{\boldsymbol{\theta}}_{\mathcal{N}}$ is finite (exists), it is unique. This implies that if Condition (b) is satisfied, Lemma 3 immediately applies to ensure the posterior of $p(\boldsymbol{\theta}_{\mathcal{N}}|\boldsymbol{y})$ is proper if $\boldsymbol{\theta}_{\mathcal{C}}$ is fixed. By the logic above, since there is a proper prior on $\boldsymbol{\theta}_{\mathcal{C}}$, the entire posterior on $p(\boldsymbol{\theta}|\boldsymbol{y})$ is proper and thus the posterior on $\boldsymbol{\beta}$ is proper.

Thus, since (a) and (b) are jointly sufficient, they are jointly necessary and sufficient as (a) alone is necessary.

A.3 Proof of Theorems 3 and 4

Theorem 3 can be established by extending results (e.g. Park and Casella 2008; Kyung et al. 2010). They note the following identity:

$$\int_0^\infty \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(\frac{-z^2}{2\tau^2} - \frac{\lambda^2\tau^2}{2}\right) \frac{\lambda^2}{2} d\tau^2 = \frac{\lambda}{2}\exp(-\lambda|z|) \tag{17a}$$

$$z \sim N(0,\tau^2); \quad \tau^2 \sim \text{Exp}(\lambda^2/2) \tag{17b}$$

The first line is crucial for our purposes as it provides a way to rewrite each K and L penalty; Equation 17 applies it to each term to get the joint density in Theorem 3. Assume a proper structured sparse prior with normalizing constant $\lambda^p c$.

$$p\left(\boldsymbol{\beta}, \{\tau_k^2\}_{k=1}^K, \{\xi_\ell^2\}_{\ell=1}^L | \lambda\right) = \lambda^p c \times \begin{array}{l} \displaystyle\prod_{k=1}^K \frac{2}{\lambda} \cdot \frac{1}{\sqrt{2\pi\tau_k^2}} \exp\left(-\frac{\boldsymbol{\beta}^T \boldsymbol{d}_k \boldsymbol{d}_k^T \boldsymbol{\beta}}{2\tau_k^2} - \frac{\lambda^2 \tau_k^2}{2}\right) \cdot \lambda^2/2 \times \\ \displaystyle\prod_{\ell=1}^L \frac{2}{\lambda} \cdot \frac{1}{\sqrt{2\pi\xi_\ell^2}} \exp\left(-\frac{\boldsymbol{\beta}^T \boldsymbol{F}_\ell \boldsymbol{\beta}}{2\xi_\ell^2} - \frac{\xi_\ell^2 \lambda^2}{2}\right) \cdot \lambda^2/2 \end{array} \tag{18}$$

Ignoring constants that do not depend on the parameters gives the result in Theorem 3. Note that the marginal prior this implies on $\{\tau_k^2\}$ and $\{\xi_\ell^2\}$ will not be the simple independent product of Gamma random variables outside of very special choices of $\boldsymbol{D}$ and $\boldsymbol{F}_\ell$ that are used in prior research (e.g. Park and Casella 2008; Kyung et al. 2010) and thus working from the joint prior is required to sample $\boldsymbol{\beta}$.

The Gibbs Sampler follows by a change of variables. Consider a single τ_k^2. The full conditional is proportional to the following; applying a change of variables gives a density that is Inverse Gaussian. The density of the Inverse Gaussian comes from Park and Casella (2008) where $\mu, \lambda > 0$.

$$p(\tau_k^2|-) \propto (\tau_k^2)^{-1/2} \exp\left(-\frac{\boldsymbol{\beta}^T \boldsymbol{d}_k \boldsymbol{d}_k^T \boldsymbol{\beta}}{2\tau_k^2} - \frac{\lambda^2 \tau_k^2}{2}\right) \tag{19a}$$

$$p(1/\tau_k^2|-) \propto (\tau_k^2)^{-3/2} \exp\left(-\lambda^2/2(1/\tau_k^2)^{-1} - \frac{1}{2}\left[\boldsymbol{\beta}^T \boldsymbol{d}_k\right]^2 \cdot (1/\tau_k^2)\right) \tag{19b}$$

$$1/\tau_k^2 \sim \text{InvGaussian}\left(\frac{\lambda}{|\boldsymbol{d}_k^T \boldsymbol{\beta}|}, \lambda^2\right) \tag{19c}$$

$$x \sim \text{InvGaussian}(\mu, \lambda) \quad \text{iff} \quad p(x|\mu,\lambda) = \sqrt{\frac{\lambda}{2\pi}} x^{-3/2} \exp\left(-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right) \tag{19d}$$

Thus, the full conditionals on all of the augmentation variables $\{\tau_k^2\}$ and $\{\xi_\ell^2\}$ are conditional independent given $\boldsymbol{\beta}$ (and λ) and all have Inverse Gaussian densities as implied by the above equation and stated in Theorem 3.

Finally, note that λ can be sampled as well. As is common in this literature, a Gamma prior is placed on λ^2 as it is conditionally conjugate with the joint density (Park and Casella 2008; Kyung et al. 2010). For some proper Gamma prior (a_0, b_0) (shape-rate parameterization), the full conditional on λ^2 is shown below

$$\lambda^2 | - \sim \text{Gamma}\left(a_0 + \frac{p+K+L}{2}, b_0 + \frac{1}{2}\left[\sum_{k=1}^K \tau_k^2 + \sum_{\ell=1}^L \xi_\ell^2\right]\right) \tag{20}$$

A.4 Extension to Global-Local Priors

The above discussion used a particular form of penalization (ℓ_1 and ℓ_2 norms) to create sparse estimates. A large literature has developed alternative Bayesian methods based on mixtures of normal distributions known as global-local priors (Polson and J. G. Scott 2011). This includes popular methods such as the adaptive LASSO, horseshoe, and others. Theorem A.1 shows that the above results are not specific to LASSO-type penalties.

Theorem A.1 *Application to Global-Local Priors*

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

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

First, define the following generalization to a multivariate $\boldsymbol{\delta}$ as follows:

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

Further, define a global-local structured sparse prior as having the following density:

$$p\left(\boldsymbol{\beta}, \{\tau_k^2\}_{k=1}^K, \{\xi_\ell^2\}_{\ell=1}^L | \boldsymbol{\lambda}\right) \propto \begin{array}{c} \exp\left(-\frac{1}{2}\boldsymbol{\beta}^T \left[\sum_{k=1}^K \frac{\boldsymbol{d}_k\boldsymbol{d}_k^T}{\tau_k^2} + \sum_{\ell=1}^L \frac{\boldsymbol{F}_\ell}{\xi_\ell^2}\right]\boldsymbol{\beta}\right) \times \\ \prod_{k=1}^K \frac{g(\tau_k^2; \boldsymbol{\lambda})}{(\tau_k^2)^{1/2}} \prod_{\ell=1}^L \frac{g(\xi_\ell^2; \boldsymbol{\lambda})}{(\xi_\ell^2)^{p/2}} \end{array} \tag{21}$$

Theorems 1 and 2 characterize the marginal prior on $\boldsymbol{\beta}$.

The proof can be established in two parts. First, I generalize Equation 22 to the case of a positive semi-definite $\boldsymbol{F}$ matrix, noting that if it is not positive definite, the prior is improper over $\boldsymbol{\delta}$.

$$p_{g,\boldsymbol{\lambda}}(\boldsymbol{\delta}) = \int_0^\infty (2\pi\tau^2)^{-p/2} \exp\left(-\frac{\boldsymbol{\delta}^T\boldsymbol{\delta}}{2\tau^2}\right) g(\tau^2; \boldsymbol{\lambda}) d\tau^2 \quad \boldsymbol{\delta}|\tau^2 \sim N(\boldsymbol{0}, \boldsymbol{I}_p \cdot \tau^2) \; \tau^2 \sim g(\tau^2; \boldsymbol{\lambda}) \tag{22}$$

For notational clarity, if I suppress $\boldsymbol{F}$ from the notation it is assumed to be the identity matrix. The product of this kernel and the analogous kernel for a linear restriction gives the proposed joint density in Theorem 3.

$$p_{g,\boldsymbol{\lambda},\boldsymbol{F}}(\boldsymbol{\delta}) = \int_0^\infty (2\pi\tau^2)^{-p/2} \exp\left(-\frac{\boldsymbol{\delta}^T\boldsymbol{F}\boldsymbol{\delta}}{2\tau^2}\right) g(\tau^2; \boldsymbol{\lambda}) d\tau^2 \tag{23}$$

With this in hand, the marginal structured sparse global-local prior on $\boldsymbol{\beta}$ has the following kernel if one integrates away the $\{\tau_k^2\}$ and $\{\xi_\ell^2\}$. Equation 24b follows by noting that since $\boldsymbol{F}_\ell$ is positive semi-definite, it can be replaced with an identity matrix.

$$
\begin{aligned}
k_{g,\boldsymbol{\lambda}}(\boldsymbol{\beta}) &= \prod_{k=1}^{K} p_{g,\boldsymbol{\lambda}}\left(\boldsymbol{d}_k^T\boldsymbol{\beta}\right) \cdot \prod_{\ell=1}^{L} p_{g,\boldsymbol{\lambda},\boldsymbol{F}}\left(\boldsymbol{\beta}\right) && \text{(24a)}\\
&= \prod_{k=1}^{K} p_{g,\boldsymbol{\lambda}}\left(\boldsymbol{d}_k^T\boldsymbol{\beta}\right) \cdot \prod_{\ell=1}^{L} p_{g,\boldsymbol{\lambda}}\left(\tilde{\boldsymbol{Q}}_\ell^T\boldsymbol{\beta}\right) && \text{(24b)}\\
&= f(\check{\boldsymbol{D}}\boldsymbol{\beta}) && \text{(24c)}
\end{aligned}
$$

Consider Theorem 1: Using the notation in the earlier proofs, assume that $\text{rank}\left(\check{\boldsymbol{D}}\right) \neq p$. The prior is improper by the same logic as before; if one defines $\boldsymbol{\theta} = \check{\boldsymbol{V}}^T\boldsymbol{\beta}$, i.e. the orthogonal rotation by the right singular matrix from a SVD of $\check{\boldsymbol{D}}$, there again are some components that appear nowhere in the prior. Thus, the integral of the kernel in Equation 24 diverges.

Sufficiency must be proven differently from above as the ℓ_1 and ℓ_2 bounds are no longer permissible. It can be shown in the following way: Assume $\check{\boldsymbol{D}}$ has full column rank. It can be expressed as follows $\check{\boldsymbol{D}} = \check{\boldsymbol{U}}_1\boldsymbol{\Sigma}\check{\boldsymbol{V}}^T$ where the full column rank means that even for the "thin" SVD, $\check{\boldsymbol{V}}^T$ is invertible. As before, I thus orthogonally rotate $\boldsymbol{\beta}$ to define $\boldsymbol{\theta}' = \boldsymbol{\Sigma}\check{\boldsymbol{V}}^T\boldsymbol{\beta}$.

$$
\begin{aligned}
\int_{\mathbb{R}^p} f(\check{\boldsymbol{D}}\boldsymbol{\beta})d\boldsymbol{\beta} &= \det(\boldsymbol{\Sigma})^{-1}\int_{\mathbb{R}^p} f(\check{\boldsymbol{U}}_1\boldsymbol{\theta}')d\boldsymbol{\theta}' && \text{(25a)}\\
&= \det(\boldsymbol{\Sigma})^{-1}\int_{\boldsymbol{z}\in\mathcal{C}(\check{\boldsymbol{D}})} f(\boldsymbol{z})d\boldsymbol{z} && \text{(25b)}\\
&\leq \det(\boldsymbol{\Sigma})^{-1}\int_{\mathbb{R}^{K+p\cdot L}} f(\boldsymbol{z})d\boldsymbol{z} < \infty && \text{(25c)}
\end{aligned}
$$

This manipulation moves from the left to the right-hand side of Equation 25a. The following line notes that $\check{\boldsymbol{U}}_1\boldsymbol{\theta}'$ is a vector of length $\mathbb{R}^{K+p\times L}$. Given that $\check{\boldsymbol{U}}_1$ is a basis for the column space of $\check{\boldsymbol{D}}$ (denoted by $\mathcal{C}(\check{\boldsymbol{D}})$), this is thus equivalent to an integral over a particular subspace of $\mathbb{R}^{K+p\times L}$.

Equation 25c follows by noting that since $f(\boldsymbol{z})$ is non-negative since it is the product of probability density functions, the integral in column space must be weakly smaller than the integral over the entire $\mathbb{R}^{K+p\times L}$ space. That integral over the entire space, however, is simply the product of (proper) probability density functions and thus is finite.

Thus, as in the LASSO case, $\check{\boldsymbol{D}}$ being full rank is necessary and sufficient for posterior propriety. Thus, by the same logic as before, examining the rank of $\bar{\boldsymbol{D}}$ characterizes the prior propriety of a global-local structured sparse prior. Note, however, that the claim following Theorem 1 about the characterization of the normalizing constant does not necessarily apply.

Equation 25b has another interesting interpretation. One can think of a structured sparse prior as "similar" to the product-of-independent sparsity inducing priors *but* that the values are constrained to lie in the column space of $\check{\boldsymbol{D}}$ to deal with the linear constraints imposed by the fact that the elements of $\boldsymbol{z}$ are not allowed to freely vary.

The final point is to show that the Theorem 2 applies. Fortunately, nothing in that proof was specific to the ℓ_1/ℓ_2 bounds except using the results of Theorem 1 and thus it follows automatically.

B Calibrating λ

There is a question of how to choose the strength of the prior (λ) for both the Bayesian and non-Bayesian approaches. This can be done in a variety of ways; cross-validation is a popular option but requires fitting the model repeatedly and thus may be computationally expensive. It further requires a "simple" data structure that can be easily partitioned into separate folds.

An alternative strategy uses information criterion such as the AIC or BIC. That requires evaluating the log-likelihood as well as a measure of complexity of the model. Tibshirani and Taylor (2012) provide an unbiased measure of the degrees of freedom for the generalized LASSO (arbitrary $\boldsymbol{D}$; $L = 0$) in the linear model that can be used to calculate these information criteria. For the non-linear case, a common approach is to use the same criterion.

In the fully Bayesian setting, one often prefers to set a prior on λ and sample it alongside $\boldsymbol{\beta}$. Appendix A shows that this can be easily done with a standard conditionally conjugate prior on λ^2 (Park and Casella 2008). Calibrating the prior, however, raises similar questions to the non-Bayesian case.

Section 6 uses a hybrid strategy where the fast posterior mode algorithm and AIC is used to find a plausible λ to anchor the prior. Specifically, I use the fast EM algorithm to perform a grid search over λ and choose the model with the best AIC. Given the optimal λ^* from this grid search, I place a conditionally conjugate prior on λ^2 (see Appendix A) where the mean and median are $(\lambda^*)^2$. As I show, the results are very similar if λ is frozen at λ^* in the Bayesian analysis.

C Details of Inference

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

C.1 Linear Regression

For linear regression, we need to incorporate the error variance σ^2 into the model. Assume the following generative framework following Park and Casella (2008) where $\text{rank}(\bar{\boldsymbol{D}}) = m$:

$$\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2 \sim N\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_N\right) \tag{26a}$$

$$p(\boldsymbol{\beta}|\lambda^2, \sigma^2) \propto w_{\bar{\boldsymbol{D}}}\lambda^m/\sigma^m \exp\left(-\frac{\lambda}{\sigma}\left[||\boldsymbol{D}\boldsymbol{\beta}||_1 + \sum_{\ell=1}^{L}\sqrt{\boldsymbol{\beta}^T\boldsymbol{F}_\ell\boldsymbol{\beta}}\right]\right) \tag{26b}$$

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

$$\begin{aligned}\ln p(\boldsymbol{\beta}|\lambda^2, \sigma^2) \propto -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2}{2\sigma^2} + \\ m\ln(\lambda) - m/2\ln(\sigma^2) - \frac{\lambda}{\sigma}\left[||\boldsymbol{D}\boldsymbol{\beta}||_1 + \sum_{\ell=1}^{L}\sqrt{\boldsymbol{\beta}^T\boldsymbol{F}_\ell\boldsymbol{\beta}}\right] + \ln p_0(\sigma^2) + \ln p_0(\lambda^2)\end{aligned} \tag{27}$$

The joint prior on $\boldsymbol{\beta}, \sigma^2, \{\tau_k^2\}, \{\xi_\ell^2\}$ follows:

$$\begin{aligned}p(\boldsymbol{\beta}, \sigma^2, \{\tau_k^2\}, \{\xi_\ell^2\}|\lambda) \propto \exp\left(-\frac{1}{2\sigma^2}\cdot\boldsymbol{\beta}^T\left[\sum_{k=1}^{K}\frac{\boldsymbol{d}_k\boldsymbol{d}_k^T}{\tau_k^2} + \sum_{\ell=1}^{L}\frac{\boldsymbol{F}_\ell}{\xi_\ell^2}\right]\boldsymbol{\beta}\right)\times \\ (\sigma^2)^{-m/2}p_0(\sigma^2)\cdot\lambda^{K+L+m}\prod_{k=1}^{K}\frac{\exp(-\lambda^2/2\cdot\tau_k^2)}{\sqrt{\tau_k^2}}\prod_{\ell=1}^{L}\frac{\exp(-\lambda^2/2\cdot\xi_\ell^2)}{\sqrt{\xi_\ell^2}}\end{aligned} \tag{28}$$

From this, the full conditional for σ^2 becomes, assuming a conjugate prior of $p_0(\sigma^2) \sim \text{InverseGamma}(a_0, b_0)$, $p_0(\lambda^2) \sim \text{Gamma}(a_{0,\boldsymbol{\Lambda}}, b_{0,\boldsymbol{\Lambda}})$ and $m = rank(\bar{\boldsymbol{D}})$.

$$\sigma^2|- \sim \text{InverseGamma}\left(a_{0,\sigma} + \frac{1}{2}\left[N + m\right], b_{0,\sigma} + \frac{1}{2}\left[\begin{array}{c}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \\ \boldsymbol{\beta}^T\left[\sum_{k=1}^{K}\frac{\boldsymbol{d}_k\boldsymbol{d}_k^T}{\tau_k^2} + \sum_{\ell=1}^{L}\frac{\boldsymbol{F}_\ell}{\xi_\ell^2}\right]\boldsymbol{\beta}\end{array}\right]\right) \tag{29}$$

The full conditionals on the other parameters are easily derived.

$$\boldsymbol{\beta}|- \sim N\left(\boldsymbol{\Sigma}_\beta\boldsymbol{X}^T\boldsymbol{y}, \sigma^2\boldsymbol{\Sigma}_\beta\right); \quad \boldsymbol{\Sigma}_\beta = \left[\boldsymbol{X}^T\boldsymbol{X} + \sum_{k=1}^{K}\frac{\boldsymbol{d}_k\boldsymbol{d}_k^T}{\tau_k^2} + \sum_{\ell=1}^{L}\frac{\boldsymbol{F}_\ell}{\xi_\ell^2}\right]^{-1} \tag{30a}$$

$$1/\tau_k^2|- \sim \text{InvGaussian}\left(\frac{\lambda\sigma}{|\boldsymbol{d}_k^T\boldsymbol{\beta}|}, \lambda^2\right) \quad 1/\xi_\ell^2|- \sim \text{InvGaussian}\left(\frac{\lambda\sigma}{\sqrt{\boldsymbol{\beta}^T\boldsymbol{F}_\ell\boldsymbol{\beta}}}, \lambda^2\right) \tag{30b}$$

$$\lambda^2|- \sim \text{Gamma}\left(a_{0,\Lambda} + [K + L + m]/2, \quad b_{0,\Lambda} + \frac{1}{2}\sum_{k=1}^{K}\tau_k^2 + \frac{1}{2}\sum_{\ell=1}^{L}\xi_\ell^2\right) \tag{30c}$$

C.2 Multinomial Regression

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

$$p(y_i = c|\{\boldsymbol{\beta}_c\}) \propto \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}_c) \tag{31}$$

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

$$\prod_{i=1}^{N} \left[\frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_c)}{\sum_{l=1}^{C} \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_l)} \right]^{I(y_i = c)} \tag{32}$$

Structured sparsity, as before, can be encoded by placing priors on $\boldsymbol{\beta}_c$. I focus on the case of identical structures for each $\boldsymbol{\beta}_c$, but one could impose more complex restrictions by constraining coefficients across-levels c in theory. The prior has the following form:

$$p(\{\boldsymbol{\beta}_c\}) \propto \prod_{c=1}^{C-1} \lambda^m \exp\left(-\lambda \left[||\boldsymbol{D}\boldsymbol{\beta}_c|| + \sum_{\ell=1}^{L} \sqrt{\boldsymbol{\beta}^T \boldsymbol{F}_\ell \boldsymbol{\beta}} \right] \right) \tag{33}$$

A Gibbs Sampler can be constructed following Polson, Scott, and Windle (2013). The key idea is to cycle through c and perform inference conditional on all other $\boldsymbol{\beta}_c$:

$$p(\{\boldsymbol{\beta}_c\}|\{\boldsymbol{\beta}_{\neg c}\}) = \prod_{i=1}^{N} \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_c - O_{ic})^{I(y_i = c)}}{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_c - O_{ic}) + 1} \cdot p(\boldsymbol{\beta}_c); O_{ic} = \ln \left(\sum_{l \neq c} \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_l) \right) \tag{34}$$

For each c, one can perform Polya-Gamma augmentation as outlined in Polson, Scott, and Windle (2013). The core identity is that, for $\omega \sim PG(1, x)$ where PG is a Polya-Gamma random variable—a particular infinite convolution of Gamman random variables:[3]

$$\boldsymbol{\beta}_c|\{\boldsymbol{\beta}_{\neg c}\} \propto \prod_{i=1}^{N} \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_c - C_{ic})^{I(y_i = c)}}{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_c - O_{ic}) + 1} \times p(\boldsymbol{\beta}_c) \tag{35a}$$

$$\omega_{i,c}|\boldsymbol{\beta}_c, \{\boldsymbol{\beta}_{\neg c}\} \sim PG(1, \boldsymbol{x}_i^T \boldsymbol{\beta}_c - O_{ic}) \tag{35b}$$

$$\begin{aligned} \boldsymbol{\beta}_c|\{\omega_{i,c}\}, \{\boldsymbol{\beta}_{\neg c}\} \sim & N\left(\boldsymbol{\Lambda}_\beta^{-1} \boldsymbol{X}^T \boldsymbol{s}, \quad \boldsymbol{\Lambda}_\beta^{-1}\right) \quad \boldsymbol{\Lambda}_\beta = \left[\sum_i \omega_{i,c} \boldsymbol{x}_i \boldsymbol{x}_i^T \right] \\ & s_i = I(y_i = c) - 1/2 - \omega_i(\boldsymbol{x}_i^T \boldsymbol{\beta}_c - O_{ic}); \quad [\boldsymbol{s}]_i = s_i \end{aligned} \tag{35c}$$

This manipulation occurs independently of the data augmentation for the sparsity penalty. Thus, one can sample the $\{\tau_k^2\}$ as before and thus create a posterior on $\boldsymbol{\beta}_c$ as follows

$$\boldsymbol{\beta}_c|\{\omega_{i,c}\}, \{\boldsymbol{\beta}_{\neg c}\}, \{\tau_k^2\}, \{\xi_\ell^2\} \sim N\left(\boldsymbol{\Sigma}_\beta \boldsymbol{X}^T \boldsymbol{s}, \boldsymbol{\Sigma}_\beta\right); \boldsymbol{\Sigma}_\beta = \left[\boldsymbol{\Lambda}_\beta + \sum_{k=1}^{K} \frac{\boldsymbol{d}_k \boldsymbol{d}_k^T}{\tau_k^2} + \sum_{\ell=1}^{L} \frac{\boldsymbol{F}_\ell}{\xi_\ell^2} \right]^{-1} \tag{36}$$

[3] Specifically, a Polya-Gamma variable is defined as below; see Polson, Scott, and Windle (2013) for details.

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

C.3 EM Algorithm

The EM algorithm follows automatically from the above results. The E-Step $(\tau_k^2; \omega_{i,c})$ is tractable and the M-Step (expectation of log complete posterior) is a simple ridge regression. It iterates these until convergence (e.g. stationarity of log-posterior or parameters).

There is one subtle point: If $\boldsymbol{d}_k^T \boldsymbol{\beta} = 0$ or $\boldsymbol{\beta}^T \boldsymbol{F} \boldsymbol{\beta} = 0$, then the corresponding augmentation variable τ_k^2 or ξ_ℓ^2 no longer has a proper density. Thus, one must deal with the fact that as $\boldsymbol{d}_k^T \boldsymbol{\beta} \to 0$, $E[1/\tau_k^2] \to \infty$ which may cause numerical instability in the algorithm. Polson and S. L. Scott (2011) suggest that when a restriction nearly binds (e.g. $|\boldsymbol{d}_k^T \boldsymbol{\beta}| < 10^{-6}$), it should be treated as binding for the remaining iterations, i.e. require that $\boldsymbol{d}_k^T \boldsymbol{\beta} = 0$ in each subsequent iteration.

I follow this logic but note their strategy relies on restricted least squares that cannot be applied to arbitrary structures. Thus, I adapt an older strategy from Lawson and Hanson (1974) discussed above and perform (unrestricted) inference in the nullspace of binding restrictions and then back-out the corresponding $\boldsymbol{\beta}$.

To avoid restrictions binding "early" by mistake or by random change, the default setting in the accompanying software is to "clip" $E[1/\tau_k^2]$ at some large value (e.g. 10^6) for the first few iterations. Further, the algorithm is initialized such that no restrictions are binding.

C.4 Adaptive LASSO

I sometimes rely on an adaptive LASSO in the spirit of Gertheiss and Tutz (2010) where the penalty is up-weighted for particular restrictions based on a consistent estimator of the difference. I also normalize each strength by the weights suggested in Gertheiss and Tutz (2010) to account for variables of different size. Both changes result in only a slight modification of the above. Assuming all of the weights are positive, this is equivalent to left multiplying $\boldsymbol{D}$ by an invertible diagonal matrix and thus all theoretical results apply automatically.

D Details on Simulations

This section outlines more detailed results on the simulations in Section 5. First, to ensure comparability, all treatment effects shown are calculated using Monte Carlo integration. I estimate the effect of d_i (moving from zero to one) for each unit g, marginalizing over x_i using Monte Carlo integration. I draw 1,000 observations from a standard normal distribution: $\{\tilde{x}_i\}_{i=1}^{1000}$. For each group g, I calculate the estimated effect for each observation i: $E[y_i | d_i = 1, g[i] = g] - E[y_i | d_i = 0, g[i] = g]$. Averaging those together gets the estimate for each method and group. This allows for comparability across all proposed models and outcomes.

Next, I enumerates the methods used with reference to the specific R packages and formulae.

- SSp - Structured Sparsity. This model is estimated using an agnostic (fully connected) structure with a LASSO penalty. I fit models over an equally spaced grid of λ on the logarithmic scale. I choose the best model using the AIC with the degrees of freedom measure in Tibshirani and Taylor (2012). Weights from Gertheiss and Tutz (2010) are used. I control for group-level effects in the simulations with a random effect estimated by approximate variational inference.

- A-SSp - Adaptive Structured Sparsity. The above model is fit with adaptive weights scaling each K restriction using a ridge-stabilized estimate of the consistent model.

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

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

- BayesGLM - From `arm`; a generalized linear model with a prior on each coefficient from Gelman et al. 2008 to avoid separation. The formula is `bayesglm(y ~ x + d * g)`

- LASSO - λ chosen using 10-fold cross-validation from `glmnet` (Friedman, Hastie, and Tibshirani 2010); the formula is `cv.glmnet(y ~ x + d * g)`. It is weighted such that there is no penalization on the main treatment effect and thus a model with maximial sparsity recovers a generalized linear model with x_i and d_i included additively.

- ENet1 - Elastic Net with $\alpha = 0.50$. λ chosen using 10-fold cross-validation from `glmnet`. Same formula as LASSO.

- ENet2 - Elastic Net with $\alpha = 0.25$. λ chosen using 10-fold cross-validation from `glmnet`. Same formula as LASSO.

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

- SVM - Estimated using polynomial kernel and package `e1071` (Meyer 2019). Default settings were used. Note that a different package, Hornik, Buchta, and Zeileis (2009), is used for the ensemble analysis.

- BART - Estimated using default settings from `BART` package (Sparapani, Spanbauer, and McCulloch 2019).

- RF - Estimated using a forest of 1,000 trees with $\lfloor G/3 \rfloor + 2$ variables drawn per tree. Estimated using the `randomForest` package (Liaw and Wiener 2002). The standard design matrix (i.e. model matrix applied to `y ~ X + treated + cgroup`) is provided.

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

Next, I discuss tree-based methods and how they include categorical predictors. Rather than including the group identifiers as indicator variables—as in all other models, some random forest approaches (e.g. Liaw and Wiener 2002) adopt different strategies. For example, one approach is to order the categories at each split based on the observed outcome and use that *ordered* variable to partition the groups (see, e.g., Hastie, Tibshirani, and Friedman 2009, p. 310).

There are two concerns with this approach; first, theoretically, it is unclear whether this is appropriate for an inferential (i.e. non-predictive) task as it, in some sense, uses the outcome to create variables to help predict the outcome! It is thus, in some sense, an "unfair" model to compare against. Second, existing software (Liaw and Wiener 2002) cannot include categorical variables with an arbitrary number of levels with certain non-standard ways of treating those variables. I thus report results from the "fair" random forest where a design matrix of indicator variables is provided.

Preliminary experiments showed that alternative methods of including the categorical predictor resulted in better performance. I replicated simulations with a linear outcome and mostly grouped predictors where I ordered the factor based on the observed response before giving it to the estimation algorithm. This resulted in markedly better performance for the random forest, although it still was handedly beaten by LASSO, random effects and structured sparsity. Exploring this in detail across different choices of software and non-standard ways of treating categorical variables is reserved for future research.

Third, Figure D.1 shows results by varying G and r. To create interpretable results, I show the percent difference in RMSE (averaged across simulations) over adaptive Structured Sparsity (A-SSp): $(\text{RMSE}_k - \text{RMSE}_{\text{A}-\text{SSp}}) / \text{RMSE}_{\text{A}-\text{SSp}} \cdot 100$. If the difference is statistically distinguishable at the 95% level, the cell is *solid*, otherwise it is light shaded.

Focusing on mostly grouped effects, when $G > 5$, structured sparsity always the best performing method and is almost always statistically distinguishably better. Note the margin of improvement is rather large; the RMSE of alternative methods is often at least 50% or over 100% worse. Similarly, it out-performs competitors even when the truth is sparse in most settings when $G > 5$. It is beaten for small r (e.g. $r = 10$, $r = 20$) by random effects, FindIt and the non-adaptive SSp in a distinguishable way, although for large r it remains the dominant method. It is worth noting that the margin of improvement is noticeably smaller, although still considerable—often around 20-50%.

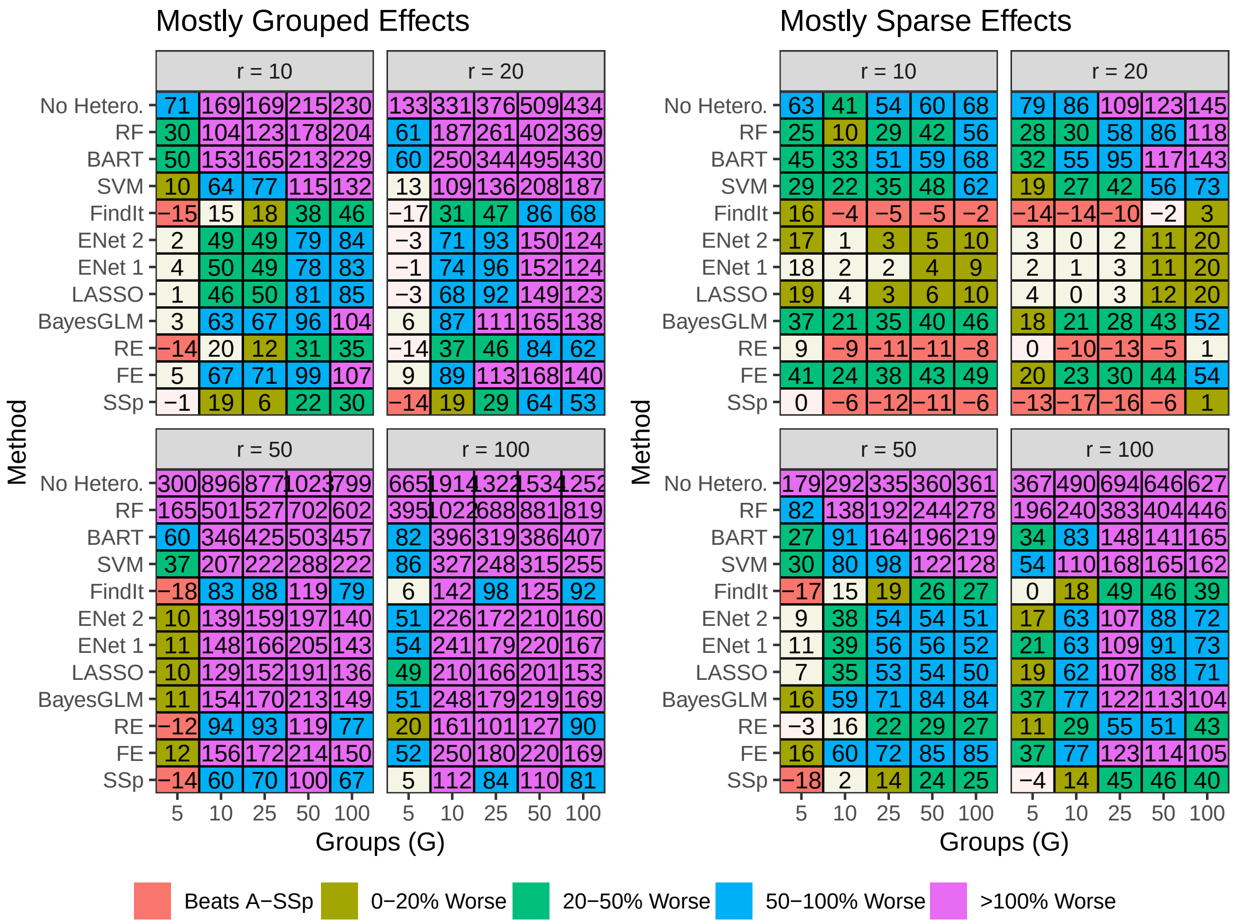
Finally, for the models with a binary outcome, the results are shown in Figure D.2. The generative model is as follows:

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

For the mostly grouped case ($\mathcal{S} = \lfloor G/2 \rfloor$), structured sparsity again out-performs almost all methods when $G > 5$. Note, however, that the magnitude of improvement is usually smaller than in the linear case although almost always statistically distinguishable; the magnitude of the improvement grows as r grows. In the mostly sparse case, the results are more mixed for structured sparsity especially when $r = 10$, but the improvements are still usually distinguishable from competitor methods although the magnitude is smaller than in the mostly grouped case—as in the linear model.

Consider an alternative summarization: Of the 80 simulation environments examined, adaptive structured sparsity performs the best in terms of mean RMSE across simulations in 43 environments and non-adaptive performs the best in 20 environments. The next best method (FindIt) performs best in 7, usually when $G = 5$. Adaptive and non-adaptive structured sparsity are in the top three methods in 60 and 74, respectively, of the environments with the next closest method (random effects) at 47.

Figure D.1: Simulations for All Methods: Linear Data Generating Process



Mostly Grouped Effects, r = 10

Method	5	10	25	50	100
No Hetero.	71	169	169	215	230
RF	30	104	123	178	204
BART	50	153	165	213	229
SVM	10	64	77	115	132
FindIt	−15	15	18	38	46
ENet 2	2	49	49	79	84
ENet 1	4	50	49	78	83
LASSO	1	46	50	81	85
BayesGLM	3	63	67	96	104
RE	−14	20	12	31	35
FE	5	67	71	99	107
SSp	−1	19	6	22	30

Mostly Grouped Effects, r = 20

Method	5	10	25	50	100
No Hetero.	133	331	376	509	434
RF	61	187	261	402	369
BART	60	250	344	495	430
SVM	13	109	136	208	187
FindIt	−17	31	47	86	68
ENet 2	−3	71	93	150	124
ENet 1	−1	74	96	152	124
LASSO	−3	68	92	149	123
BayesGLM	6	87	111	165	138
RE	−14	37	46	84	62
FE	9	89	113	168	140
SSp	−14	19	29	64	53

Mostly Grouped Effects, r = 50

Method	5	10	25	50	100
No Hetero.	300	896	877	1023	799
RF	165	501	527	702	602
BART	60	346	425	503	457
SVM	37	207	222	288	222
FindIt	−18	83	88	119	79
ENet 2	10	139	159	197	140
ENet 1	11	148	166	205	143
LASSO	10	129	152	191	136
BayesGLM	11	154	170	213	149
RE	−12	94	93	119	77
FE	12	156	172	214	150
SSp	−14	60	70	100	67

Mostly Grouped Effects, r = 100

Method	5	10	25	50	100
No Hetero.	665	1914	1322	1534	1252
RF	395	1022	688	881	819
BART	82	396	319	386	407
SVM	86	327	248	315	255
FindIt	6	142	98	125	92
ENet 2	51	226	172	210	160
ENet 1	54	241	179	220	167
LASSO	49	210	166	201	153
BayesGLM	51	248	179	219	169
RE	20	161	101	127	90
FE	52	250	180	220	169
SSp	5	112	84	110	81

Mostly Sparse Effects, r = 10

Method	5	10	25	50	100
No Hetero.	63	41	54	60	68
RF	25	10	29	42	56
BART	45	33	51	59	68
SVM	29	22	35	48	62
FindIt	16	−4	−5	−5	−2
ENet 2	17	1	3	5	10
ENet 1	18	2	2	4	9
LASSO	19	4	3	6	10
BayesGLM	37	21	35	40	46
RE	9	−9	−11	−11	−8
FE	41	24	38	43	49
SSp	0	−6	−12	−11	−6

Mostly Sparse Effects, r = 20

Method	5	10	25	50	100
No Hetero.	79	86	109	123	145
RF	28	30	58	86	118
BART	32	55	95	117	143
SVM	19	27	42	56	73
FindIt	−14	−14	−10	−2	3
ENet 2	3	0	2	11	20
ENet 1	2	1	3	11	20
LASSO	4	0	3	12	20
BayesGLM	18	21	28	43	52
RE	0	−10	−13	−5	1
FE	20	23	30	44	54
SSp	−13	−17	−16	−6	1

Mostly Sparse Effects, r = 50

Method	5	10	25	50	100
No Hetero.	179	292	335	360	361
RF	82	138	192	244	278
BART	27	91	164	196	219
SVM	30	80	98	122	128
FindIt	−17	15	19	26	27
ENet 2	9	38	54	54	51
ENet 1	11	39	56	56	52
LASSO	7	35	53	54	50
BayesGLM	16	59	71	84	84
RE	−3	16	22	29	27
FE	16	60	72	85	85
SSp	−18	2	14	24	25

Mostly Sparse Effects, r = 100

Method	5	10	25	50	100
No Hetero.	367	490	694	646	627
RF	196	240	383	404	446
BART	34	83	148	141	165
SVM	54	110	168	165	162
FindIt	0	18	49	46	39
ENet 2	17	63	107	88	72
ENet 1	21	63	109	91	73
LASSO	19	62	107	88	71
BayesGLM	37	77	122	113	104
RE	11	29	55	51	43
FE	37	77	123	114	105
SSp	−4	14	45	46	40

Note: The percentage change in RMSE vs A-SSp (adaptive structured sparsity) is shown: $\left(\text{RMSE}_k - \text{RMSE}_{\text{A-SSp}}\right)/\text{RMSE}_{\text{A-SSp}} \cdot 100$. A positive number thus indicates *worse* performance. All blocks that are lightly shaded indicate a difference that is *not* statistically distinguishable at the 95% level. The number reported is averaged across 100 simulations. For example, with $r = 10$, mostly grouped effects, $G = 5$, A-SSp beats an SVM by 10% and loses to FindIt by 15% (although the latter is not distinguishable from zero).

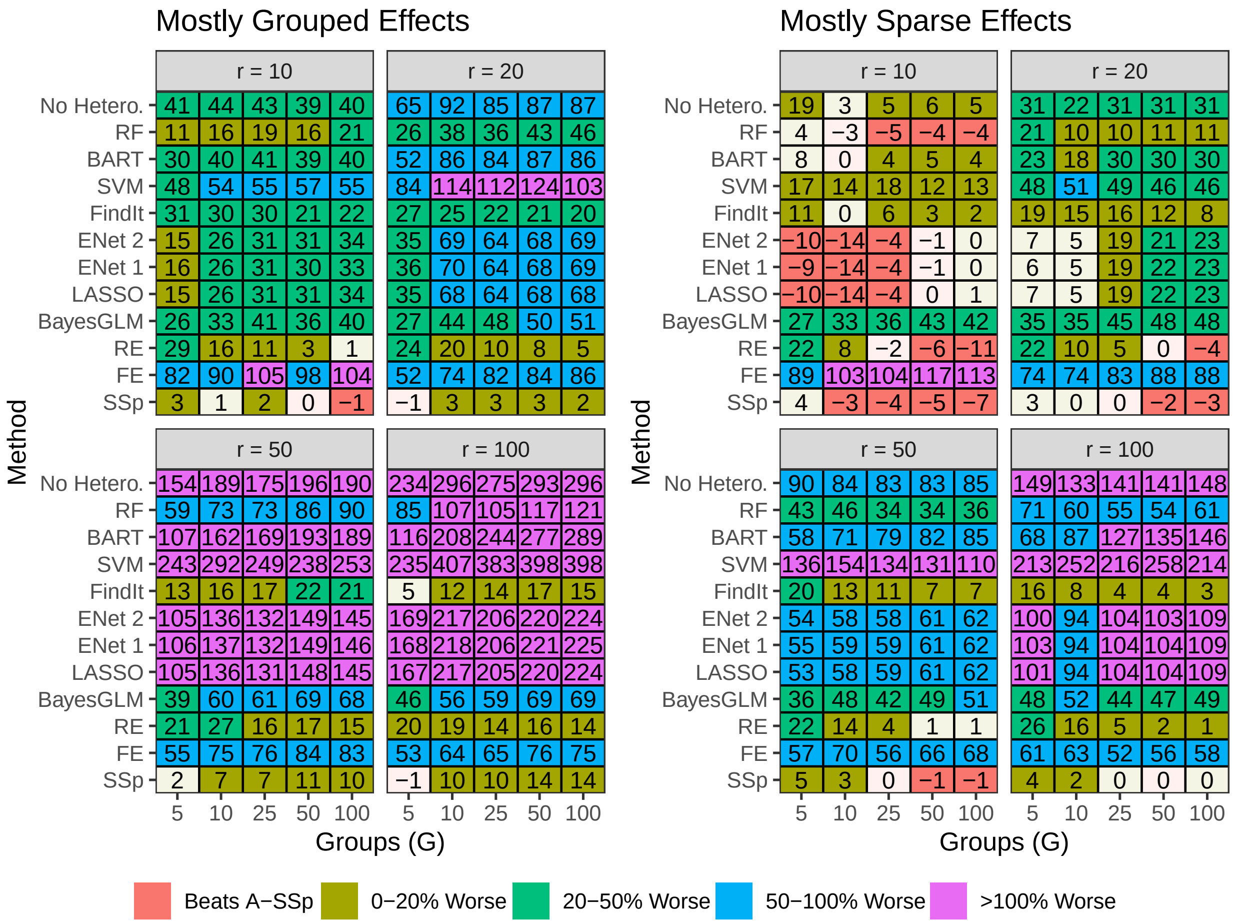
A final remark is that non-adaptive structured sparsity appears to out-perform structured sparsity when the truth is mostly sparse (wining 25/40 times vs. 9/40).

E Details on Credit-Claiming Analysis

E.1 Bayesian Convergence Diagnostics

I present results on the convergence of the posterior sampler for the three structured sparse models outlined in the main text. I ran each model with over-dispersed starting values (drawing from a uniform ranging from -3 to 3) for 4 chains, 10,000 iterations each and discarded the first 5,000 as burn-in. This gives 20,000 samples from the posterior. Figure E.3a

Figure D.2: Simulations for All Methods: Multinomial Data Generating Process



Mostly Grouped Effects

r = 10

Method	5	10	25	50	100
No Hetero.	41	44	43	39	40
RF	11	16	19	16	21
BART	30	40	41	39	40
SVM	48	54	55	57	55
FindIt	31	30	30	21	22
ENet 2	15	26	31	31	34
ENet 1	16	26	31	30	33
LASSO	15	26	31	31	34
BayesGLM	26	33	41	36	40
RE	29	16	11	3	1
FE	82	90	105	98	104
SSp	3	1	2	0	−1

r = 20

Method	5	10	25	50	100
No Hetero.	65	92	85	87	87
RF	26	38	36	43	46
BART	52	86	84	87	86
SVM	84	114	112	124	103
FindIt	27	25	22	21	20
ENet 2	35	69	64	68	69
ENet 1	36	70	64	68	69
LASSO	35	68	64	68	68
BayesGLM	27	44	48	50	51
RE	24	20	10	8	5
FE	52	74	82	84	86
SSp	−1	3	3	3	2

r = 50

Method	5	10	25	50	100
No Hetero.	154	189	175	196	190
RF	59	73	73	86	90
BART	107	162	169	193	189
SVM	243	292	249	238	253
FindIt	13	16	17	22	21
ENet 2	105	136	132	149	145
ENet 1	106	137	132	149	146
LASSO	105	136	131	148	145
BayesGLM	39	60	61	69	68
RE	21	27	16	17	15
FE	55	75	76	84	83
SSp	2	7	7	11	10

r = 100

Method	5	10	25	50	100
No Hetero.	234	296	275	293	296
RF	85	107	105	117	121
BART	116	208	244	277	289
SVM	235	407	383	398	398
FindIt	5	12	14	17	15
ENet 2	169	217	206	220	224
ENet 1	168	218	206	221	225
LASSO	167	217	205	220	224
BayesGLM	46	56	59	69	69
RE	20	19	14	16	14
FE	53	64	65	76	75
SSp	−1	10	10	14	14

Mostly Sparse Effects

r = 10

Method	5	10	25	50	100
No Hetero.	19	3	5	6	5
RF	4	−3	−5	−4	−4
BART	8	0	4	5	4
SVM	17	14	18	12	13
FindIt	11	0	6	3	2
ENet 2	−10	−14	−4	−1	0
ENet 1	−9	−14	−4	−1	0
LASSO	−10	−14	−4	0	1
BayesGLM	27	33	36	43	42
RE	22	8	−2	−6	−11
FE	89	103	104	117	113
SSp	4	−3	−4	−5	−7

r = 20

Method	5	10	25	50	100
No Hetero.	31	22	31	31	31
RF	21	10	10	11	11
BART	23	18	30	30	30
SVM	48	51	49	46	46
FindIt	19	15	16	12	8
ENet 2	7	5	19	21	23
ENet 1	6	5	19	22	23
LASSO	7	5	19	22	23
BayesGLM	35	35	45	48	48
RE	22	10	5	0	−4
FE	74	74	83	88	88
SSp	3	0	0	−2	−3

r = 50

Method	5	10	25	50	100
No Hetero.	90	84	83	83	85
RF	43	46	34	34	36
BART	58	71	79	82	85
SVM	136	154	134	131	110
FindIt	20	13	11	7	7
ENet 2	54	58	58	61	62
ENet 1	55	59	59	61	62
LASSO	53	58	59	61	62
BayesGLM	36	48	42	49	51
RE	22	14	4	1	1
FE	57	70	56	66	68
SSp	5	3	0	−1	−1

r = 100

Method	5	10	25	50	100
No Hetero.	149	133	141	141	148
RF	71	60	55	54	61
BART	68	87	127	135	146
SVM	213	252	216	258	214
FindIt	16	8	4	4	3
ENet 2	100	94	104	103	109
ENet 1	103	94	104	104	109
LASSO	101	94	104	104	109
BayesGLM	48	52	44	47	49
RE	26	16	5	2	1
FE	61	63	52	56	58
SSp	4	2	0	0	0

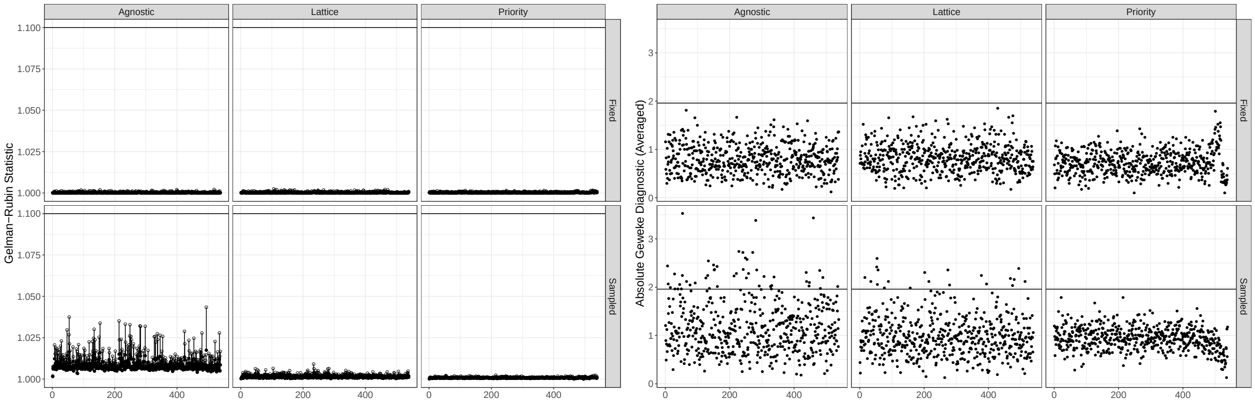
Note: The percentage change in RMSE vs A-SSp (adaptive structured sparsity) is shown: $(\text{RMSE}_k - \text{RMSE}_{\text{A-SSp}})/\text{RMSE}_{\text{A-SSp}} \cdot 100$. A positive number thus indicates *worse* performance. All blocks that are lightly shaded indicate a difference that is *not* statistically distinguishable at the 95% level. The number reported is averaged across 100 simulations. For example, with $r = 10$, mostly grouped effects, $G = 5$, A-SSp beats an SVM by 48% and beats FindIt by 31%; both are distinguishable from zero.

reports the Gelman-Rubin statistic for all parameters in the model. A threshold of 1.1 is a common test for convergence. All parameters are below this value as are the upper confidence statistic reported by `coda`.

I also report the Geweke statistic. To summarize across chains, I report the average (absolute) statistic across all four chains. Looking at each statistic individually, it is above the 1.96 threshold in below 5% of parameter-chain combinations when λ is fixed. When λ is sampled, it is higher; around 12% for the lattice model used in the main text.

The likely cause of this poorer behavior is the issues with the convergence of λ when sampled for the lattice and agnostic structure. Even after considerable burn-in, mixing appears to be slow as Figure E.4 shows—especially for the agnostic structure. A different sampler may perform better. Exploring this is an important area for future research. Fortunately, as the results in the next sub-section show, fixing λ at λ^* returns nearly identical estimates

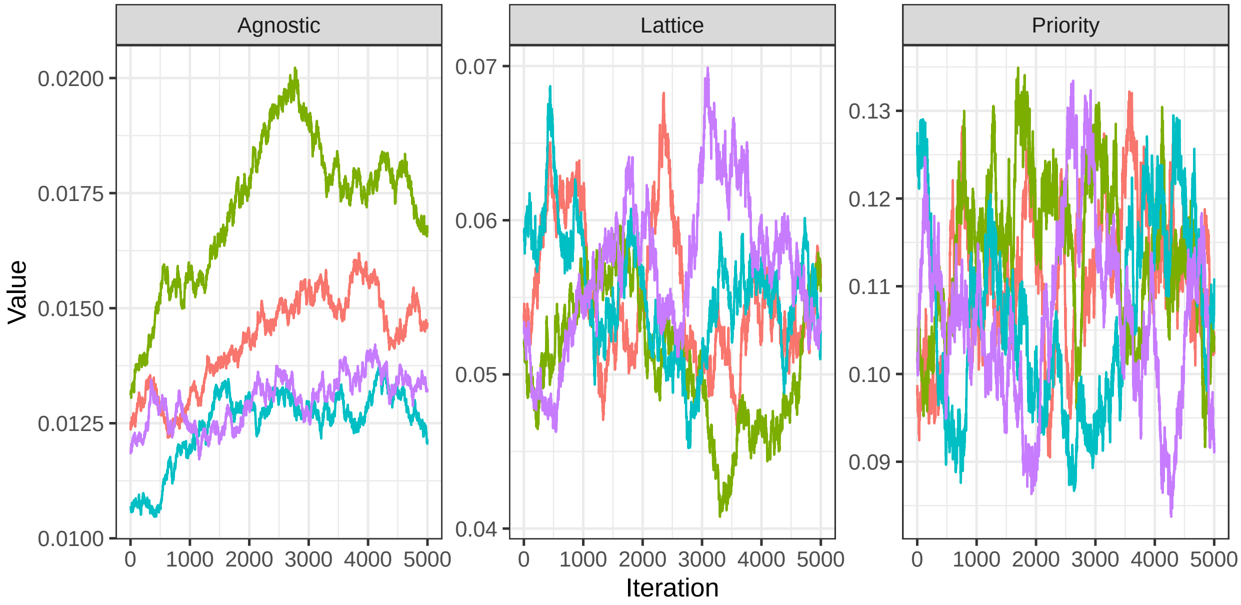
Figure E.3: Convergence Diagnostics on $\boldsymbol{\beta}$


(a) Gelman-Rubin Convergence Statistics

(b) Geweke Convergence Statistics

suggesting that the lack of stationarity in λ will not materially undermine the results.

Figure E.4: Lambda Trace Plot

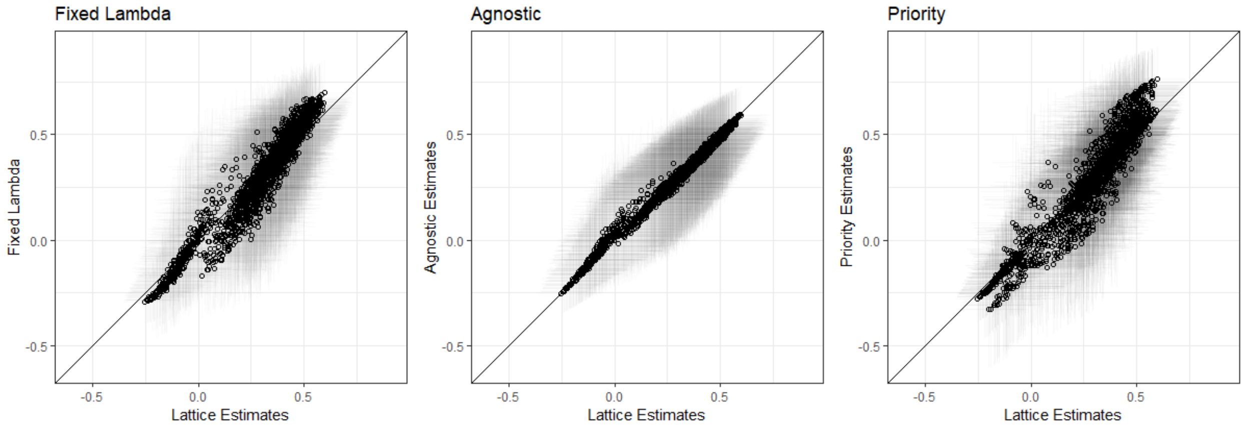


I attempted to fix this problem by running a much longer model for the lattice structure: 20,000 iterations after 20,000 of burn-in on four chains. This improves within-chain mixing (i.e. the Geweke diagnostic for all chains is below 1.96), but the Gelman-Rubin statistic is still stubbornly high at 1.17.

E.2 Substantive Additional Results

Figure E.5 shows the stability of treatment effects across structure. It plots the estimated treatment effects for all treatment/respondent combinations. The lattice effects (reported in the main text) are shown on the horizontal axis. Each panel reports a particular comparison; the model estimated with a fixed λ^* at the optimal value found via the AIC, an agnostic structure and a priority structure.

Figure E.5: Comparison of Structures



Note: Each figure plots the posterior median of one of three alterantive methods (lattice structure with fixed λ, agnostic structure, and priority structure) against the estimates with a lattice structure. The 90% credible intervals for each parameter are shown in light grey below.

Note the tight correlation between the estimates with a fixed λ and lattice structure and allowing it to be sampled. This provides confidence that the results are not being driven by the lack of stationarity in the sampling of λ. The comparison with the agnostic method also suggests that the choice of structure may be less important if the limiting conditions are the same. The priority results look somewhat different but are closely correlated.

References

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48.

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.

Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. 2008. "A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models." *The Annals of Applied Statistics* 2 (4): 1360–1383.

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

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

Hornik, Kurt, Christian Buchta, and Achim Zeileis. 2009. “Open-Source Machine Learning: R Meets Weka.” *Computational Statistics* 24 (2): 225–232.

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

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

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

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

Meyer, David. 2019. “Support Vector Machines: The Interface to libsvm in package e1071.” Accessed on Aug 21 2019. https://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf.

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

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

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

Polson, Nicholas G., 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. 2011. “Data Augmentation for Support Vector Machines.” *Bayesian Analysis* 6 (1): 1–24.

Sparapani, Rodney, Charles Spanbauer, and Robert McCulloch. 2019. “The BART R package.” Accessed on Aug 21 2019. https://rdrr.io/cran/BART/f/inst/doc/the-BART-R-package.pdf.

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

Tibshirani, Ryan J., and Jonathan Taylor. 2012. “Degrees of Freedom in Lasso Problems.” *The Annals of Statistics* 40 (2): 1198–1232.