A Variable-Coefficient Nuclear Norm Penalty for Low Rank Inference

author names withheld

Under Review for OPT 2022

Abstract

Low rank structure is expected in many applications, so it is often desirable to be able to specify cost functions that induce low rank. A common approach is to augment the cost with a penalty function approximating the rank function, such as the nuclear norm which is given by the ℓ_1 norm of the matrix's singular values. This has the advantage of being a convex function, but it biases matrix entries towards zero. On the other hand, nonconvex approximations to the rank function can make better surrogates but invariably introduce additional hyperparameters. In this article, we instead study a weighted nuclear norm approach with learnable weights which provides the behavior of nonconvex penalties without introducing any additional hyperparameters. This approach can also benefit from the fast proximal methods which make nuclear norm approaches scalable. We demonstrate the potential of this technique by comparing it against the standard nuclear norm approach on synthetic and realistic matrix denoising and completion problems. We also outline the future work necessary to deploy this algorithm to large scale problems.

1. Introduction

Oftentimes it is desirable for a matrix-valued optimization quantity to have low rank. The function $\text{rank}(\mathbf{A})$ of $\mathbf{A} \in \mathbb{R}^{M\times N}$ unfortunately is difficult to optimize directly, being nonconvex and discontinuous. Therefore, the nuclear or Ky Fan norm [8] $\|\mathbf{A}\|_*$, defined as the sum of the $\mathbf{A}$'s singular values, is often used as a surrogate loss function as it is convex and continuous, leading to the following optimization problem:

$$\min_{\mathbf{A}\in\mathbb{R}^{M\times N}} f(\mathbf{A}) + \tau\|\mathbf{A}\|_* = \min_{\mathbf{A}\in\mathbb{R}^{M\times N}} f(\mathbf{A}) + \tau\sum_{p=1}^{P}\sigma_p(\mathbf{A}) \ ,$$

where $\sigma_p(\mathbf{A})$ gives the p'th singular value of $\mathbf{A}$, and we $P = \min(M, N)$. This class of problems can be solved efficiently for smooth f using proximal gradient methods [15]. Most saliently, Cai et al. [2], Mazumder et al. [13] deployed the Iterative Shrinkage and Thresholding Algorithm [4, ISTA] with f given by the matrix completion problem (see Section 4).

In this article, we propose using a version of the nuclear norm with variable weights, to be estiamted via optimization, which will allow for a better surrogate loss function in the sense that it more closely mimicks the behavior of the ℓ_0 norm while maintaining a continuous optimization problem. We begin this article in Section 2 by providing an overview of pertinent sparsity and low-rank inducing penalties. In Section 4, we deploy our novel penalty on noisy matrix completion problem based on natural images. Finally, Section 5 concludes and overviews future research on this topic and related ones.

© .

2. Background

In the vector case, ISTA is parameterized by a step size η and regularization strength τ and proceeds by the proximal gradient formula, which involves iteration of two steps:

$$\begin{aligned}\mathbf{x}_0^{t+1} &\leftarrow \mathbf{x}^t - \eta \nabla f(\mathbf{x}^t) \\ \mathbf{x}^{t+1} &\leftarrow \mathrm{STO}_{\eta\tau}(\mathbf{x}_0^{t+1}) \ ,\end{aligned}$$

where STO is the Soft Thresholding Operator applied elementwise: $\mathrm{STO}_{\eta\tau}(x) = (|x| - \eta\tau)^+ \mathrm{sgn}(x)$. In the matrix case, iteration proceeds by applying the STO elementwise to the singular values of the solution matrix after the gradient descent step:

$$\begin{aligned}&\mathbf{A}_0^{t+1} \leftarrow \mathbf{A}^t - \eta \nabla f(\mathbf{A}^t) \\ &\mathbf{U}^{t+1}, \boldsymbol{\sigma}^{t+1}, \mathbf{V}^{t+1,\top} \leftarrow \mathrm{svd}(\mathbf{A}_0^{t+1}) \\ &\mathbf{A}^{t+1} \leftarrow \mathbf{U}^{t+1} \mathrm{diag}[\mathrm{STO}_{\eta\tau}(\boldsymbol{\sigma}^{t+1})] \mathbf{V}^{t+1,\top} \ ,\end{aligned}$$

where $\mathrm{svd}(\mathbf{A})$ is the singular value decomposition of $\mathbf{A}$.

The nuclear norm serves as a convex relaxation of the rank function [9] much as the ℓ_1 norm serves as a convex relaxation of the ℓ_0 "norm" for vectors. ℓ_1 penalized regression [20] is now ubiquitous in many domains of computational science, and is called *Lasso regression* [21] in the machine learning community, having been catapulted to popularity in part by its ease of computation and interpretability. However, this penalty does have some drawbacks when viewed as a surrogate of the ℓ_0 function, particularly when there are large values among the nonzero signal. The practical consequence of this is bias towards zero: in the matrix case, imposition of a nuclear norm penalty means that though some components will be successfully thresholded to zero, the nonzero components will be overly shrunk towards zero, sometimes perniciously so.

For this reason, a slew of debiased penalties have been proposed to replace the ℓ_1 norm with a better surrogate of the ℓ_0 norm. Many of these are nonconvex penalties originally proposed for sparsity in regression problems such as as the Minimax Concave Penalty [23, MCP], the Smoothly Clipped Absolute Deviation [7, SCAD] and Bridge [10] penalties. [17, 22] develop a general optimization framework that can accommodate many nonconvex penalties. Marjanovic and Solo [12] brought Bridge penalties to the matrix case in order to perform matrix completion using the fact that the proximal operator of the ℓ_q penalty is known. [14], on the other hand, develop an iterative reweighting procedure for the bridge norm, defining a weighted version of the nuclear norm: $\|\mathbf{A}\|_*^{\boldsymbol{\lambda}} = \sum_{p=1}^P \lambda_p \sigma_p(\mathbf{A})$. [11] extend this to more nonconvex penalties. In this article, we will also make use of a weighted nuclear norm, but rather than choose λ_p in order to locally approximate a prespecified nonconvex loss, we instead propose to treat $\boldsymbol{\lambda}$ as a decision variable and to determine its value via optimization.

The *Maximum a Posteriori* (MAP) Bayesian perspective on Lasso regression [16] specifies the prior distribution $\beta \sim \mathrm{L}(0, \frac{1}{\lambda})$, where $\mathrm{L}(0, \frac{1}{\lambda})$ gives a Laplace distribution with inverse scale parameter λ. Methods in the family of the Spike-Slab Lasso [19] or Horseshoe prior [3] specify a coefficient-specific regularization parameter λ_p endowed with a common marginal hyperprior P_λ, which alleviates the bias that the standard Laplace (or Normal in the Horseshoe case) prior places on selected coefficients. Classically, such hyperpriors are then used inside of an MCMC procedure.

In this article, we will instead be interested in the MAP perspective:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^P,\boldsymbol{\lambda}\geq\mathbf{0}} f(\boldsymbol{\beta}) + \tau\sum_{p=1}^{P}\lambda_p|\beta_p| - \sum_{p=1}^{P}\log P_\lambda(\lambda_p)$$

Optimization of this quantity is complicated by the coupling between λ_p and β_p which renders the STO inappropriate for joint optimization. In an article under preparation which has been appended to this manuscript, we showed that there is nevertheless a closed form proximal operator appropriate for this problem. If η is a step size common to all variables and assuming that $\tau\eta^2 \leq 1$, the *Variable* Soft Thresholding Operator is given by:

$$\mathrm{VSTO}_{\tau\eta}(\lambda_0, x_0): \quad \lambda^* = \begin{cases} \lambda_0 & \lambda_0 \geq \frac{|x_0|}{\tau\eta} \\ \frac{(\lambda_0-\tau\eta|x_0|)^+}{1-\tau^2\eta^2} & o.w. \end{cases}, \tag{1}$$

$$x^* = (|x_0| - \eta\lambda^*)^+\mathrm{sgn}(x_0) \tag{2}$$

where $(a)^+ = \max(0, a)$, and $\lambda^* = \mathbb{1}_{[\lambda_0>|x_0|]}\lambda_0$ when $\tau\eta^2 > 1$ (with x^* unchanged). We refer to the deployment of the VSTO proximal operator and priors on $\boldsymbol{\lambda}$ within a gradient descent procedure as Variable ISTA, or VISTA. In this article, we extend this approach to the matrix case to allow for adaptive low-rank inference.

3. A Variable-Coefficient Nuclear Norm Penalty

We propose to endow $\mathbf{A}$ with the prior matrix distribution with density $\delta(\mathbf{A}) \propto e^{-\|\mathbf{A}\|_*^{\boldsymbol{\lambda}}}$, and then to specify independent hyperpriors P_λ for each λ_p, then to conduct MAP inference with this prior structure. The optimization problem thus becomes:

$$\min_{\mathbf{A}\in\mathbb{R}^{M\times N};\boldsymbol{\lambda}\geq\mathbf{0}} f(\mathbf{A}) + \tau\sum_{p=1}^{P}\lambda_p\sigma_p(\mathbf{A}) - \sum_{p=1}^{P}\log P_\lambda(\lambda_p)\ .$$

Iteration proceeds with a gradient descent step with respect to $\mathbf{A}$ and $\boldsymbol{\lambda}$ before applying the VSTO operator to $\boldsymbol{\lambda}$ and the singular values of $\mathbf{A}$:

$$\begin{aligned}
&\mathbf{A}_0^{t+1} \leftarrow \mathbf{A}^t - \eta\nabla f(\mathbf{A}^t) \\
&\boldsymbol{\lambda}_0^{t+1} \leftarrow \boldsymbol{\lambda}^t - \eta\nabla\big[-\sum_{p=1}^{P}\log P_\lambda(\boldsymbol{\lambda}^t)\big] \\
&\mathbf{U}^{t+1}, \boldsymbol{\sigma}_0^{t+1}, \mathbf{V}^{t+1,\top} \leftarrow \mathrm{svd}(\mathbf{A}_0^{t+1}) \\
&\boldsymbol{\lambda}^{t+1}, \boldsymbol{\sigma}^{t+1} \leftarrow \mathrm{VSTO}_{\tau\eta}(\boldsymbol{\lambda}_0^{t+1}, \boldsymbol{\sigma}_0^{t+1}) \\
&\mathbf{A}^{t+1} \leftarrow \mathbf{U}^{t+1}\mathrm{diag}(\boldsymbol{\sigma}^{t+1})\mathbf{V}^{t+1,\top}\ .
\end{aligned}$$

Unlike the previously proposed nonconvex penalties enumerated by [11], the variable-coefficient nuclear norm penalty does not introduce any additional tuning parameters. Though it relies instead on the specification of a hyperprior P_λ, we have found the standard Half Cauchy prior to be appropriate for the applications in this article as well as in the regression problems in the draft manuscript attached to this article.

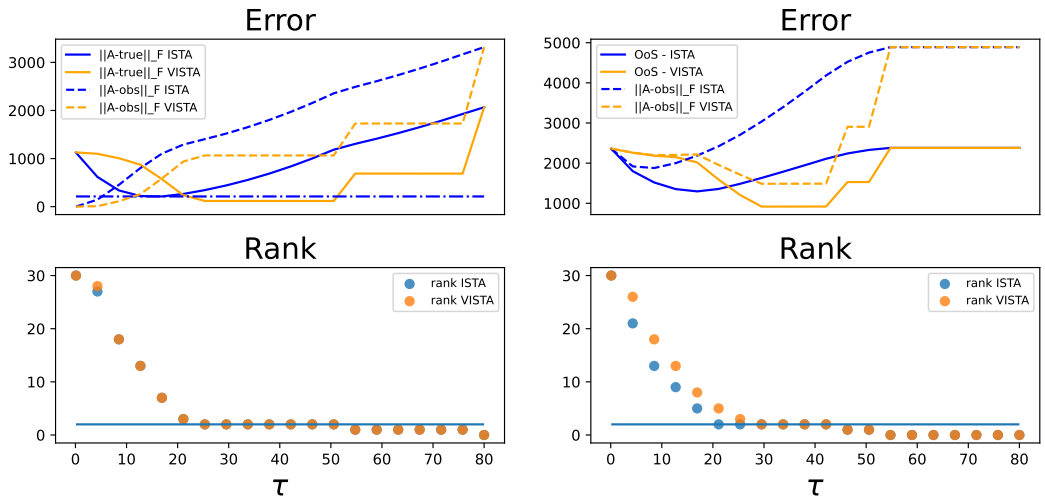

Figure 1: **Synthetic Illustration on Individual Matrices:** $\mathbf{A} \in \mathbb{R}^{30\times 40}$ with standard normal noise. *Left:* Denoising. *Right:* Noisy Matrix Completion. Dotted lines give reconstruction error, solid lines give prediction error.

One difference between the vector and matrix cases is the fact that the singular values come with a natural ordering, whereas a vector's elements comes only with a nominal ordering. It might be thus desired to have $\lambda_p \leq \lambda_{p+1}$. But this would couple the optimization problems, significantly complicating proximal operator computation. In this article, we simply apply the prox to each singular value individually, and observe that $\lambda_p \leq \lambda_{p+1}$ naturally. Since the singular values themselves are of course ordered, we would expect that this would impose the correct orderings on λ_p in most cases, but have not yet established a proof for all P_λ nor developed a counterexample.

4. Applications

In this section, we first qualitatively compare the proposed VISTA approach with the classical ISTA on two random matrices with before quantitatively comparing the methods on completion of noisy natural images. Neither VISTA nor ISTA are competitive with modern approaches for natural image completion that can exploit the spatial layout of the pixels such as wavelets or convolutional neural networks. We use this example nevertheless as images are a classic benchmark for matrix completion and we are able to use images small enough so as to be able to explicitly compute singular values decompositions.

First, we consider the low rank denoising problem, that is, we generate a matrix $\mathbf{A}_{\text{obs}} = \mathbf{B} + \mathbf{E}$, where $\mathbf{B}$ is of rank 2 and $\mathbf{E}$ is a matrix of standard normal random variates. The classical cost function for this problem is:

$$\min_{\mathbf{A}\in\mathbb{R}^{M\times N}} \|\mathbf{A}_{\text{obs}} - \mathbf{A}\|_F^2 + \tau \sum_{p=1}^{P} \sigma_p(\mathbf{A})] \ ,$$

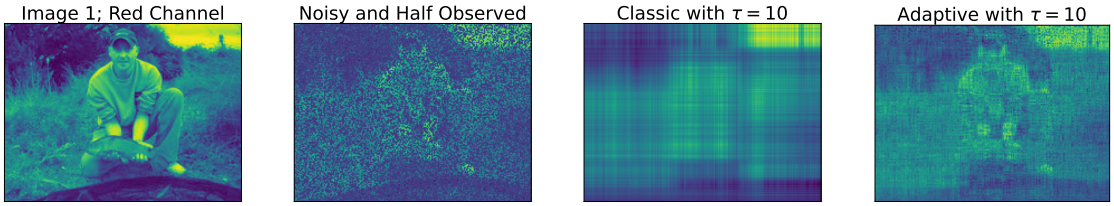

Figure 2: **Example Case:** *Left to Right:* Original Image, Noisy partial image presented to algorithms, ISTA solution, VISTA solution.

which has a closed form in terms of the singular value decomposition of $\mathbf{A}_{\text{obs}} = \mathbf{U}\text{diag}[\sigma]\mathbf{V}^\top$ [2]:

$$\mathbf{A}^* = \mathbf{U}\text{diag}[(\boldsymbol{\sigma} - \tau\mathbf{1})^+]\mathbf{V}^\top \ .$$

The cost function associated with the adaptive nuclear penalty is instead:

$$\min_{\mathbf{A}\in\mathbb{R}^{M\times N};\boldsymbol{\lambda}\geq\mathbf{0}} \|\mathbf{A}_{\text{obs}} - \mathbf{A}\|_F^2 + \tau\sum_{p=1}^{P}\lambda_p\sigma_p(\mathbf{A}) - \sum_{p=1}^{P}\log P_\lambda(\lambda_p)$$

which we solve iteratively.

We also consider the matrix completion problem, which is similar except for we only have data constraints for a subset of $\mathbf{A}_{\text{obs}}$'s entries, which we encode as an observation mask $m_{i,j}$ which takes value 1 if the entry is observed and 0 otherwise:

$$f(\mathbf{A}) = \|(\mathbf{A}_{\text{obs}} - \mathbf{A}) \odot \mathbf{M}\|_F^2$$

again with both penalties. Unlike the denoising case, this cost function augmented with the classical Nuclear penalty does not enjoy a closed form solution, and a common solution algorithm is ISTA. See [5] for more background and algorithms.

We deploy both ISTA and VISTA as part of a gradient descent algorithm with fixed step size of 10^{-3} and for 2,000 iterations.

4.1. Illustrative Synthetic Examples

Figure 1 shows the trajectory of the reconstruction error as τ varies. In the left figure, all entries of the matrix are observed with noise, with error evaluated by distance to the noiseless low-rank matrix (i.e. $\mathbf{B}$). We notice that the rank of the solution for a given τ is identical between VISTA and ISTA. However, the approximation errors are quite different. The ISTA algorithm achieves its optimal error at $\tau \approx 20$, using an approximation with rank greater than 2. VISTA, on the other hand, achieves its optimal error at the correct rank of 2 and for a wide range of τ values between about 20 and 50. It would seem that ISTA is forced to let in singular values that should be zero in order to avoid excessive shrinkage of the truly nonzero components. VISTA, on the other hand, is able to identify a good solution with the correct rank.

The right side of the figure shows a similar setup but now with only 50% of the entries available. Error is now evaluated with respect to left-out matrix entries. Again, VISTA is able to find a lower

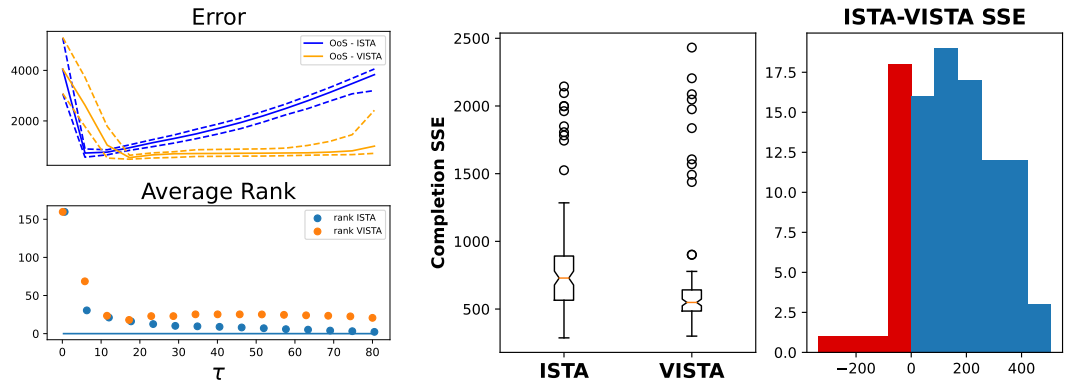

Figure 3: **Imagenet Test Problems:** *Left:* The plot gives the 2.5, 50, and 97.5 percentiles of sum squared prediction error for both algorithms for a range of τ. *Center:* The distribution of out of sample SSE for VISTA is below that of ISTA. *Right:* A histogram of the SSE differences on the considered images; a value greater than 0 indicates that VISTA is performing better on that image (such bars are colored blue).

rank solution which achieves better Out of Sample error (i.e. error on the unseen matrix entries) than ISTA is. Unlike the completely observed case, VISTA sometimes chooses different ranks for a give τ value than ISTA, tending to favor larger rank solutions at smaller τ values. Notice that the ISTA error curve varies continuously with τ, whereas VISTA has flat regions followed by sudden changes. This indicates that VISTA returns the same approximations for ranges of τ values, meaning that it is less sensitive to hyperparameter specification than is ISTA.

4.2. Small Natural Image Completion and Denoising

We now deploy adapative and classical proximal nuclear norm minimization to a natural image benchmark. We use images from Imagenet [6], in particular the imagenette[1] subset. This repo has 963 images of widths between 160 and 269 and heights between 160 and 480. We treat the R, G and B channels as individual matrices, and choose as our benchmark the 100 first such matrices. We scale them to $[0, 1]$, then add i.i.d. Gaussian noise with a standard deviation of 0.1 and sample the observation mask from a bernoulli with $p = 0.5$ such that the algorithm sees about half the image with noise and has to fill in the other half (see Figure 2, right and center-left).

Figure 3 presents the distribution of Sum Squared Error on held-out matrix entries for each algorithm. The top-left panel shows us the 25'th, 50'th, and 95'th percentiles of the error for each algorithm, revealing that ISTA has its best performance for smaller τ than VISTA. Additionally, it would seem that VISTA can achieve peak performance for a wider range of τ than can ISTA, suggesting again that it is less sensitive to specification of its global penalty parameter. The lower-left panel shows the average rank of solutions from each algorithm; they are higher for VISTA for most τ values. We next compare the error distributions associated with each method for the specific τ at which each performs best. We choose $\tau \approx 5$ for ISTA and $\tau \approx 15$ for VISTA. The middle panel

1. https://github.com/fastai/imagenette

shows the distribution of error for each as pair of boxplots while the right panel shows a histogram of pairwise error differences, revealing that VISTA does better for almost all images, though there are a significant number of images for which it does especially poorly.

Though VISTA generally gives better average performance, certain images flummoxed it, at least for the regularization level chosen. We suspect that in certain applications it would be helpful to include some amount of "non-adaptive" regularization by augmenting cost with a standard ℓ_2 or ℓ_1 regularization as suggested in the context of regression by Piironen and Vehtari [18].

5. Conclusion and Future Work

This article introduced a variable-coefficient version of the nuclear norm penalty for inducing low-rank structure in matrices. We compared it to the classic singular value thresholding approach on test problems small enough that we could actually compute the SVD to apply our proximal operator. But in many interesting applications this is not possible, notably large scale recommendation problems ("The Netflix Problem"). Authors working with the nuclear norm have been able to successfully deploy their procedure within scalable singular value estimation procedures, and we see no obstacle to doing the same with our adaptive norm; we look forward to this future work.

We caution that the adaptive nuclear norm should not be viewed as universally superior to the standard kind. In some applications, particularly those where there is not true low rank structure, the shrinkage imposed by the nuclear norm may be desired, and the bias it induces may prove stabilizing.

The denoising problem with complete observation can be solved in closed form via a single application of the classical nuclear norm proximal operator to the singular values of the observed noisy matrix. Conceivably, the same is true of our adaptive method, depending on the P_λ chosen. In future work, we will identify a suitable prior and investigate the possibility of a closed form adaptive denoiser.

Previous work has shown that Nesterov acceleration may be profitably incorporated in proximal gradient methods (notably [1]), and there is no reason to believe that this would not be the case here (though it may destroy special structure in $\mathbf{A}$ such as sparsity).

References

[1] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[2] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4):1956–1982, 2010.

[3] Carlos M Carvalho, Nicholas G Polson, and James G Scott. Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*, pages 73–80. PMLR, 2009.

[4] Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11): 1413–1457, 2004.

[5] Mark A Davenport and Justin Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):608–622, 2016.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[7] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[8] Ky Fan. Maximum properties and inequalities for the eigenvalues of completely continuous operators. *Proceedings of the National Academy of Sciences*, 37(11):760–766, 1951.

[9] Maryam Fazel, Haitham Hindi, and Stephen P Boyd. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the 2001 American Control Conference.(Cat. No. 01CH37148)*, volume 6, pages 4734–4739. IEEE, 2001.

[10] Wenjiang J Fu. Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics*, 7(3):397–416, 1998.

[11] Canyi Lu, Jinhui Tang, Shuicheng Yan, and Zhouchen Lin. Nonconvex nonsmooth low rank minimization via iteratively reweighted nuclear norm. *IEEE Transactions on Image Processing*, 25(2):829–839, 2015.

[12] Goran Marjanovic and Victor Solo. On lq optimization and matrix completion. *IEEE Transactions on signal processing*, 60(11):5714–5724, 2012.

[13] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11:2287–2322, 2010.

[14] Karthik Mohan and Maryam Fazel. Iterative reweighted algorithms for matrix rank minimization. *The Journal of Machine Learning Research*, 13(1):3441–3473, 2012.

[15] Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014.

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

[17] Duy Nhat Phan and Thuy Ngoc Nguyen. An accelerated irnn-iteratively reweighted nuclear norm algorithm for nonconvex nonsmooth low-rank minimization problems. *Journal of Computational and Applied Mathematics*, 396:113602, 2021.

[18] Juho Piironen and Aki Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2):5018–5051, 2017.

[19] Veronika Ročková and Edward I George. The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444, 2018.

[20] Howard L Taylor, Stephen C Banks, and John F McCoy. Deconvolution with the ℓ_1 norm. *Geophysics*, 44(1):39–52, 1979.

[21] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. URL http://www.jstor.org/stable/2346178.

[22] Quanming Yao, James T Kwok, Taifeng Wang, and Tie-Yan Liu. Large-scale low-rank matrix learning with nonconvex regularizers. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2628–2643, 2018.

[23] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

Appendix A. Supplementary Material

Dear Editor;
Dear Reviewer:

Please find below a draft manuscript under review which contains, among other things, the derivation of the VSTO operator and VISTA algorithm referred to in the above work. For a potential Camera Ready version of this manuscript, we intend to have submitted a preprint to arXiv which we will cite, or the citation of the accepted paper.

On Behalf of the Authors,
Anonymous

DRAFT UNDER REVIEW: Sparse Bayesian Lasso via a Biconvex Proximal Algorithm

author names withheld

September 30, 2022

Abstract

Proximal algorithms in the form of the Iterative Shrinkage and Thresholding Algorithm (ISTA) extend ℓ_1 penalized inference with exact sparsity to arbitrary smooth likelihoods, allowing for model selection via Lasso penalty in MAP inference with complicated models. However, the ℓ_1 penalty introduces significant bias when high sparsity is desired, an issue addressed by simulation-based approaches to variable selection such as the Spike Slab or Horseshoe priors which vary the prior variance terms serving as inverse penalty coefficients. We begin this article by examining the optimization problem associated with a Laplace prior with a hyperprior on its inverse scale parameter. On the way, we discuss a pertinent proximal operator and deploy it to develop a variable-penalty version of ISTA we term VISTA. We show how to use VISTA to perform debiased MAP variable selection. Next, we develop a nonsmooth variational Bayesian framework which demonstrates good empirical frequentist properties in simulation studies. This gives us the UQ and low bias properties of simulation-based approaches with an order of magnitude less computation. Finally, we demonstrate the potential of VISTA in general complex likelihoods by applying it to a Bayesian lagged spatiotemporal regression model of internal displacement that occurred during the 2015 to 2017 years of the Iraqi Civil War.

Keywords: variational Bayesian inference, sparsity, proximal operators, nonsmooth optimization, gravity model, human migration

1 Introduction

Basic Inferential Problem: Sparsifying penalties are popular because they allow for simultaneous variable selection and parameter estimation by shrinking parameter values identically to 0. Explicitly, we define the sparse learning problem as follows:

$$\min_{\boldsymbol{\beta},\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\beta},\boldsymbol{\theta}) + \tau||\boldsymbol{\beta}||_0 \ , \tag{1}$$

where $\boldsymbol{\beta}$ is a parameter vector which we wish to be sparse, $\boldsymbol{\theta}$ is a vector of other model parameters and $\mathcal{L}$ is a term which determines the compatibility of a parameter configuration with a dataset (such as a log-likelihood or log posterior density). This is a difficult problem because the $||.||_0$ function, which counts the nonzero components of its input, is nonconvex and discontinuous. The $||.||_1$ norm is a useful approximation because it is the closest convex function to $||.||_0$. This is the penalty function of the celebrated Lasso (Tibshirani, 1996; Taylor et al., 1979), which may be viewed as placing a Laplace prior on β_p (Park and Casella, 2008). The draw-back to this convexity is bias: estimates of nonzero parameters will be shrunk towards zero, significantly so if we are to achieve much sparsity (Zhang, 2010). One solution to this problem is to allow the penalty coefficients to vary by parameter:

$$\min_{\boldsymbol{\beta},\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\beta},\boldsymbol{\theta}) + \tau \sum_{p=1}^{P} \lambda_p |\beta_p| \ , \tag{2}$$

as when $\lambda_p \to 0$ the associated β_p is unpenalized. Far from clear is how to choose these λ_p, as evidenced by the assortment of approaches discussed in the next section.

Existing Approaches and Limitations: The importance of variable selection in non-linear models has prompted decades of research and yielded a multitude of approaches. In this section, we review selected penalty-based approaches to variable selection.

Some approaches perform adaptation "offline", that is, before or after the optimization process. The Adaptive Lasso (Zou, 2006) specifies $\lambda_p = \frac{1}{|\hat{\beta}_p|^\gamma}$, where $\hat{\beta}_p$ is some initial

estimate of the regression coefficient and γ is a hyperparameter, though an initial estimate $\hat{\beta}_p$ is a nontrivial ask if $\mathcal{L}$ is complicated. Bühlmann and Meier (2008); Candes et al. (2008) propose to iterate this procedure, updating the penalty coefficient λ_p with new $\hat{\beta}_p$. Alternatively, we may adopt a two-stage process, using Lasso for variable selection only and proceeding to an unpenalized procedure (Efron et al., 2004; Meinshausen, 2007; Zou and Li, 2008).

Several authors have proposed simulation-based procedures for adaptation of λ_p. Kang and Guo (2009); Leng et al. (2014); Mallick and Yi (2014) place a Gamma prior or similar for λ_p and conduct inference via Gibbs sampling. Bhattacharya et al. (2012) instead take the approach of specifying a Dirichlet prior on the regression coefficients before conducting MCMC. Conceptually this is similar to the Horseshoe Prior (Carvalho et al., 2010), which instead specifies a conditional Normal prior for β_p. The Horseshoe prior enjoys generality as well as fast specific implementations (Terenin et al., 2019; Makalic and Schmidt, 2015). The reader is referred to Bhadra et al. (2019) for an extensive comparison of the Lasso and Horseshoe models. Another approach is the Spike-Slab prior (Mitchell and Beauchamp, 1988) which uses discrete latent variables to categorize whether β_p was *a priori* sampled from the spike or the slab, which would complicate gradient-based inference.

Variational Bayesian inference (Blei et al., 2017, VB) is another way to adapt the λ_p. But there is a sense in which VB with a Laplace prior differs from Lasso: the integrated cost erases the sparsifying geometry (see Section 3.1.2). Previous work has circumvented this using either a point mass variational distribution for regression coefficients (Tung et al., 2019), which does not allow for uncertainty quantification, or by thresholding small coefficients (Babacan et al., 2014), which requires setting an arbitrary threshold (She, 2009). Kawano et al. (2015) vary the threshold while monitoring information criteria.

Nonconvex penalties, on the other hand, do not vary the regularization strength but are directly constructed to impose minimal bias on nonzero coefficients, such as the Smoothly Clipped Absolute Deviation function (Fan and Li, 2001; Hunter and Li, 2005, SCAD) or Minimax Concave Penalty (Zhang, 2010, MCP). These penalties do not correspond to proper priors, as they place constant, positive probability density arbitrarily far from the origin.

Scientific Motivation: Forced human migration is at an all time high and only increasing. In order for policymakers to plan effectively, it is important to understand which factors affect when and where people will move during times of crisis. Traditionally, migration researchers use *gravity models*, described in Section 4.1.2. These models were developed by economist to model international trade (the flow of goods and services) as a function of traditional econometric indicators, typically on a coarse spatiotemporal scale (i.e. country by year). Migration researchers use this model because it can be effective in capturing spatial interactions (both push and pull dynamics). In this article, we study an approximately year-long period of the Iraqi Civil War of 2013-2017, where the government and its international backers fought ISIL in the north of the country. Traditional migration variables are not immediately available in exactly those situations where they are needed most: dangerous, remote, and underdeveloped areas, such as our case study. Taking advantage of the proliferation of the mobile internet, we develop novel social media "Buzz Variables" derived from Arabic-Language twitter (Section 4.1.1). This allows us to make predictions at finer spatiotemporal scales than traditional country-year level administrative data, but introduces new challenges, namely unignorable lag between cause and effect, spatiotemporal correlation, and high dimensional and noisy predictors. In order to meet these new challenges, we develop a Bayesian hierarchical model which allows for zero inflation,

overdispersion, spatiotemporal random effects, and lagging and aggregation of predictors (Section 4.2) – a complex model to perform variable selection on.

Brief Outline of Our Contributions: We begin this article with a study of the approach of treating the Laplace inverse scale parameter $\boldsymbol{\lambda} := \{\lambda_1, \ldots, \lambda_P\}$ as an additional parameter to be optimized, one endowed with the hyperprior p_λ, which yields the following optimization:

$$\min_{\beta,\theta,\boldsymbol{\lambda}>\mathbf{0}} \mathcal{L}(\beta,\theta) + \sum_{p=1}^{P} \left[\tau\lambda_p|\beta_p| + \log\lambda_p\right] + \sum_{p=1}^{P} -\log p_\lambda(\lambda_p) \ , \tag{3}$$

where τ is a positive parameter controlling the sparsity level. The ℓ_1 penalty in general may be efficiently applied to a smooth loss function via the Iterative Shrinkage and Thresholding Algorithm (Daubechies et al., 2004, ISTA), a proximal gradient method Parikh et al. (2014). But ISTA assumes a known and fixed λ_p. Section 2 examines a proximal operator associated with the variable-λ_p optimization problem and discusses how to deploy it to marry adaptive and nonsmooth penalties in an analog of ISTA we term Variable ISTA (VISTA) [1].

In Section 3 we develop a Variational Bayesian take on an adaptive Lasso using a nonsmooth penalty we term the Sparse Bayesian Lasso (SBL). The "Bayesian Lasso" of Park and Casella (2008) is so named as it is a Bayesian explanation for the density used in Lasso. But the Laplace prior does not actually encode the prior beliefs that motivate the use of Lasso, i.e., that the coefficient vector is sparse. Rather, it encodes the belief that most of the coefficients are near zero, which leads to poor interval coverage and bias. As such it is not *operationally* a Bayesian Lasso, that is, it is not a method that yields the sparsity properties of Lasso as part of a Bayesian analysis. The proposed Sparse Bayesian Lasso fills this role by allowing for both geometric sparsity and uncertainty quantification. Combined with warm starts, we can efficiently calculate trajectories not just of parameter

[1] As we deploy Nesterov acceleration, the better analog is actually Fast ISTA (Beck and Teboulle, 2009).

estimates, as in the traditional Lasso, but of entire variational distributions, showing us how uncertainty of model parameters change with sparsity. This allows us to avoid specifying or estimating an overall model complexity parameter τ, which is difficult in practice.

Most practitioners using bespoke statistical models do so in a software environment that makes inference straightforward to conduct given a model specification. In this article, we explain how to deploy our method in gradient-based learning frameworks such as `tensorflow` and `pytorch`. After verifying the frequentist properties of the SBL in simulation studies, we demonstrate the potential of the SBL in complex hierarchical models using our Iraq forced displacement case study (Section 4). We conclude by discussing research directions suggested by the novel proximal operator in Section 5.

2 Coefficient-Specific λ_p with Optimization

We begin this section by providing background on proximal operators and methods. Then, we investigate the proximal operator associated with the ℓ_1 norm when the penalty coefficient is considered a decision variable before discussing how to embed it in optimization.

2.1 Background: Nonsmooth Penalties and Proximal Operators

We are interested in minimizing a complicated but smooth loss l augmented with a simple but nonsmooth regularizer g:

$$c(\mathbf{x}) = l(\mathbf{x}) + \lambda g(\mathbf{x}) \,, \tag{4}$$

where g is given by the ℓ_1 norm in the case of Lasso regression. Proximal gradient descent and its relatives are the algorithms of choice in such a situation. Given a function g with domain $\mathcal{X}$ and some norm parameterized by a positive definite matrix $\mathbf{C}$, their *proximal*

operator is then defined as the following mapping:

$$\text{prox}_g^{\mathbf{C}}(\mathbf{x}) = \underset{\mathbf{u}\in\mathcal{X}}{\text{argmin}}\, g(\mathbf{u}) + \frac{1}{2}||\mathbf{x}-\mathbf{u}||_{\mathbf{C}}^2 \ . \tag{5}$$

Intuitively, the proximal operator of a function g evaluated at a vector $\mathbf{x}$ returns another vector $\mathbf{u}$ which is close to $\mathbf{x}$ (wrt $\mathbf{C}$) but does a better job minimizing g.

Proximal Gradient algorithms optimize the objective function in Equation 4 via iteration of a two step process. Given a current solution $\mathbf{x}^k$ and smooth cost function $\mathcal{L}$, the next iterate is parameterized by a step size η and given by:

$$\hat{\mathbf{x}}^{k+1} = \mathbf{x}^k - \eta\nabla\mathcal{L}(\mathbf{x}^k) \tag{6}$$

$$\mathbf{x}^{k+1} = \text{prox}_g^{\eta\mathbf{C}}(\hat{\mathbf{x}}^{k+1})\,. \tag{7}$$

Note that $\mathbf{C}$ is scaled by the step size in the proximal operator and that we have not included a subgradient of g in the gradient descent step. The matrix $\mathbf{C}$ may in theory be any positive definite matrix and in practice it is defined by the preconditioner of the gradient descent algorithm that the proximal operator will be deployed in.

The proximal operator is most useful when it can be computed efficiently. In this article, we will assume a diagonal preconditioner $\mathbf{C} = \text{diag}(c_1, \ldots, c_P)$, which is conducive to breaking the proximal problem into subproblems defined along each axis. For example, when $g : \mathbb{R} \to \mathbb{R}$ is given by $g(x) = \lambda|x|$, the proximal operator is given by elementwise application of the Soft Thresholding Operator (STO):

$$[\text{prox}_{\lambda|.|}^{\eta\text{diag}(\mathrm{c}_1,\ldots,\mathrm{c}_\mathrm{P})}(\mathbf{x})]_p = (|x_p| - \eta c_p\lambda)^+\text{sgn}(x_p)\,, \tag{8}$$

where $(a)^+$ gives $\max(0, a)$ and $\text{sgn}(a)$ gives the sign of a.

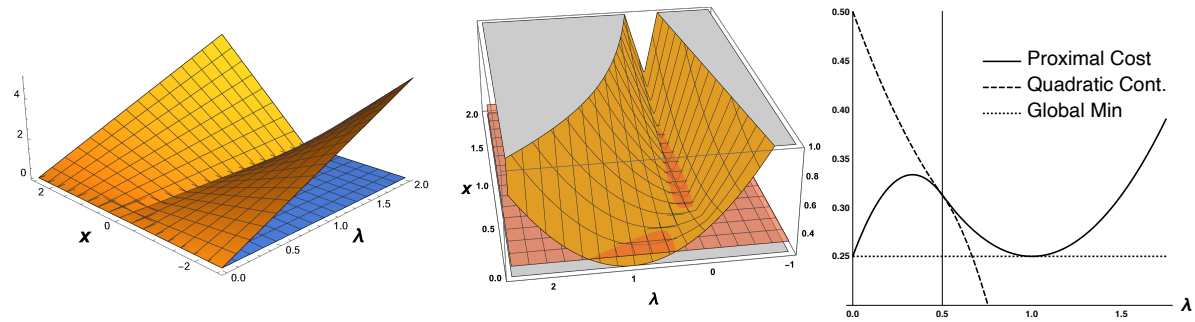

Figure 1: **The Proximal Cost** *Left:* The function $g(x, \lambda) = \lambda|x|$. The proximal cost *Center:* of (x, λ) *Right:* marginal for λ with $\lambda_0 = x_0 = 1; \eta_\lambda = \eta_x = 2$ yielding two optima.

2.2 The Variable-Coefficient ℓ_1 Proximal Operator

Because the ℓ_1 regularization coefficient λ is now being optimized over, the STO is no longer the pertinent proximal operator. Indeed, the ℓ_1 regularization function, considered formally as a function of both λ and $\mathbf{x}$, is nonconvex (see Figure 1, Left), somewhat complicating proximal operator computation. In fact, many authors such as Parikh et al. (2014) define the prox operator as one that acts on convex functions. However, there has been work on extending proximality to larger classes of functions (see e.g. (Hare and Sagastizábal, 2009)) and, of particular interest to the statistical community, development of proximal operators for the nonconvex "Bridge Penalties" $|\beta_p|^q$ for $q \in (0, 1)$ (Marjanovic and Solo, 2013). Perhaps because of this focus on convexity (and despite the popularity of the ℓ_1 norm and adaptive penalty methods), the proximal operator of $\lambda|x|$ as a $\mathbb{R}^+ \times \mathbb{R} \to \mathbb{R}^+$ function has not to our knowledge been previously examined in the literature. It turns out that the action of this proximal operator is available in closed form. It is single-valued for almost all inputs and always for sufficiently small step sizes η_x and η_λ such that $\eta_x \eta_\lambda < 1$.

Consider the proximal operator of the weighted ℓ_1 norm function $g(\mathbf{x}, \boldsymbol{\lambda}) = \sum_{p=1}^{P} \lambda_p |x_p|$. Since this function decomposes into additive functions of each $(\lambda_p,\ x_p)$ individually, its proximal operator acts on each block independently of the others. Therefore, for the

remainder of this section, we consider a single block (λ, x), dropping the index p, and compute the proximal operator of the 2-dimensional function $g(x, \lambda) = \lambda|x|$:

Problem: $$\text{prox}_g^{\eta_\lambda, \eta_x}(x_0, \lambda_0) = \underset{x \in \mathbb{R}, \lambda > 0}{\text{argmin}} \ \lambda|x| + \frac{(x - x_0)^2}{2\eta_x} + \frac{(\lambda - \lambda_0)^2}{2\eta_\lambda} \tag{9}$$

Solution: Convert to nested optimization and exploit the fact that the marginal solution for fixed x is known, with optimizing $x^* = (|x_0| - \eta_x \lambda)^+ \text{sgn}(x_0)$:

$$\underset{x \in \mathbb{R}, \lambda > 0}{\text{argmin}} \ \lambda|x| + \frac{(x - x_0)^2}{2\eta_x} + \frac{(\lambda - \lambda_0)^2}{2\eta_\lambda} \tag{10}$$

$$\iff \underset{\lambda > 0}{\text{argmin}} \ \frac{(\lambda - \lambda_0)^2}{2\eta_\lambda} + \underset{x \in \mathbb{R}}{\text{argmin}} \ \lambda|x| + \frac{(x - x_0)^2}{2\eta_x} \tag{11}$$

$$\iff \underset{\lambda > 0}{\text{argmin}} \ \frac{(\lambda - \lambda_0)^2}{2\eta_\lambda} + \lambda(|x_0| - \eta_x \lambda)^+ + \frac{(|x_0| - \eta_x \lambda)^{+,2} - 2|x_0|(|x_0| - \eta_x \lambda)^+}{2\eta_x} \ . \tag{12}$$

This may be rewritten as a piecewise polynomial with changepoint given by $\lambda = \frac{|x_0|}{\eta_x}$, the point where λ is just large enough to push x to zero:

$$\underset{\lambda > 0}{\text{argmin}} \begin{cases} \frac{1}{2}(\frac{1}{\eta_\lambda} - \eta_x)\lambda^2 + (|x_0| - \frac{\lambda_0}{\eta_\lambda})\lambda + \frac{\lambda_0^2}{2\eta_\lambda} & \lambda < \frac{|x_0|}{\eta_x} \\ \frac{(\lambda - \lambda_0)^2}{2\eta_\lambda} + \frac{x_0^2}{2\eta_x} & \lambda \geq \frac{|x_0|}{\eta_x} \ . \end{cases} \tag{13}$$

The quadratic polynomial in the interval $[\frac{|x_0|}{\eta_x}, \infty]$ is always convex. When $\eta_\lambda \eta_x < 1$, the quadratic polynomial in the other interval is convex as is the overall expression. But when $\eta_\lambda \eta_x > 1$, the coefficient of the quadratic term is negative, and that polynomial is concave, yielding a nonconvex piecewise function (see Figure 1, center and right).

We now find the optimum of each interval. When $\lambda \geq \frac{|x_0|}{\eta_x}$, the optimum is simply as close as we can get to λ_0, namely $\lambda \leftarrow \max[\lambda_0, \frac{x_0}{\eta_x}]$. On the other hand, when $\lambda \leq \frac{|x_0|}{\eta_x}$, if $\eta_x \eta_\lambda < 1$, the optimum is as close as we can get to the stationary point $\frac{(\lambda_0 - \eta_\lambda |x_0|)}{1 - \eta_\lambda \eta_x}$, explicitly $\lambda \leftarrow \min[\frac{(\lambda_0 - \eta_\lambda |x_0|)^+}{1 - \eta_\lambda \eta_x}, \frac{|x_0|}{\eta_x}]$. When $\eta_x \eta_\lambda \geq 1$, however, the solution is at one of the interval boundaries $[0, \frac{|x_0|}{\eta_x}]$; the boundaries have costs of $\frac{\lambda_0^2}{2\eta_\lambda}$ and $\frac{(\frac{|x_0|}{\eta_x} - \lambda_0)^2}{2\eta_\lambda} + \frac{x_0^2}{2\eta_x}$, respectively, and

so we choose $\lambda \leftarrow 0$ if $\frac{\lambda_0^2}{2\eta_\lambda} < \frac{(\frac{|x_0|}{\eta_x} - \lambda_0)^2}{2\eta_\lambda} + \frac{x_0^2}{2\eta_x}$ and $\lambda \leftarrow \frac{|x_0|}{\eta_x}$ otherwise. But the cost at $\lambda = \lambda_0$ is only $\frac{x_0^2}{2\eta_x}$. Therefore, the choice is between 0 and λ_0 with costs $\frac{\lambda_0^2}{2\eta_\lambda}$ and $\frac{x_0^2}{2\eta_x}$.

So when $\eta_x \eta_l < 1$, we have:

$$\lambda^* = \begin{cases} \lambda_0 & \lambda_0 \geq \frac{|x_0|}{\eta_x} \\ \frac{(\lambda_0 - \eta_\lambda |x_0|)^+}{1 - \eta_\lambda \eta_x} & o.w. \quad , \end{cases} \tag{14}$$

when $\eta_x \eta_l \geq 1$ we instead have: $\lambda^* = \mathbb{1}_{\left[\frac{\lambda_0}{\sqrt{\eta_l}} > \frac{|x_0|}{\sqrt{\eta_x}}\right]} \lambda_0$ (here $\mathbb{1}$ denotes the indicator function), and in either case $x^* = (|x_0| - \eta_x \lambda^*)^+ \mathrm{sgn}(x_0)$.

□

A result of this nonconvexity is that this proximal program may have two global optima. A consequence of this in turn is that the proximal operator is discontinuous and multi-valued at the discontinuity, as visualized in the top right of Figure 2.

Notice that when $\eta_x \eta_\lambda < 0$ and $\lambda_0 < \eta_\lambda |x_0|$ or when $\eta_x \eta_\lambda > 1$ and $\frac{\lambda_0}{\sqrt{\eta_\lambda}} > \frac{|x_0|}{\sqrt{\eta_x}}$, the solution to the proximal problem gives $\lambda = 0$, which would lead to no shrinkage on β. In our application of this operator to the Laplace density, however, the λ normalization term gives this a density of zero.

This proximal operator has been conceptualized as a mapping of $(\lambda_0, x_0) \to (\lambda^*, x^*)$ parameterized by η_x and η_λ, but to visualize it we will briefly study it as a function of these four quantities mapping to an optimizing λ^*. Since when $\eta_x \eta_\lambda > 1$ the λ^* is either 0 or λ_0, we will focus on the case where $\eta_x \eta_\lambda < 1$. Then, let $a := \frac{|x_0|}{\eta_x}$ and $b := \eta_x \eta_\lambda$. We can rewrite the solution as a function of just three variables:

$$\lambda(\lambda_0, a, b) = \begin{cases} \lambda_0 & \lambda_0 \geq a \\ \frac{(\lambda_0 - ab)^+}{1 - b} & o.w. \quad . \end{cases} \tag{15}$$

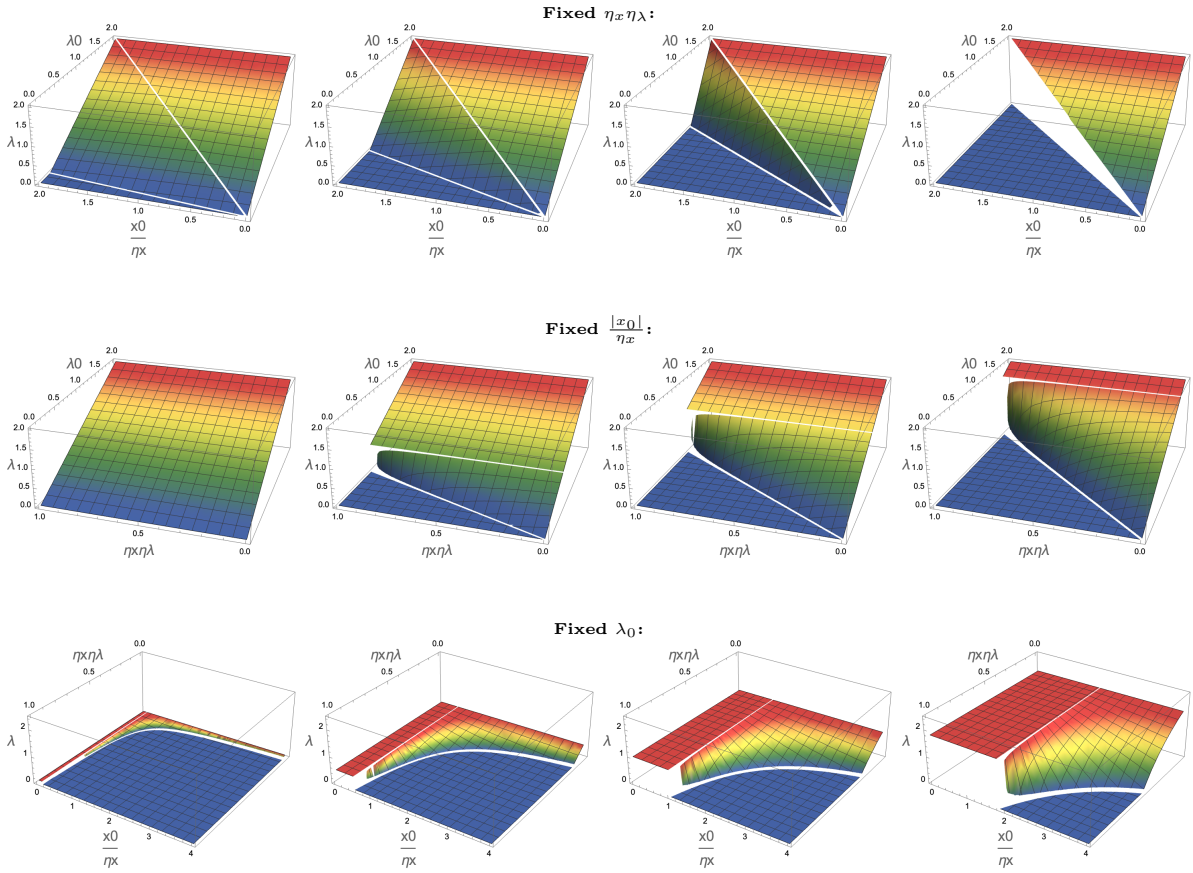

Figure 2: **The Action of the Proximal Operator:** Plots of the reduced proximal operator (Eq 15) considering two variables at a time and fixing the third. *Top:* for various fixed $b := \eta_x \eta_\lambda < 1$ and with $\lambda_0, \frac{|x_0|}{\eta_x} \in (0, 2)$. Values $b = \eta_x \eta_\lambda \in \{0.1, 0.35, 0.65, 0.99\}$ are shown left to right. *Mid:* $a = \frac{|x_0|}{\eta_x} \in \{0, 0.5, 1, 1.8\}$. *Bottom:* $\lambda_0 \in \{0.2, 0.75, 1.25, 1.75\}$.

This function is visualized in Figure 2. For given step size product b the function is stepwise linear, and converges to the identity mapping with respect to its λ_0 input as $b \to 0$. As $b \to 1$, the mapping becomes more and more steep for $\lambda_0 \in (ab, a)$, gradually converging to the discontinuous mapping $\lambda(\lambda_0, a, b) \to \lambda_0 \mathbb{1}_{\lambda_0 > a}$. The proximal operator is multi-valued at $\lambda_0 = a$ when $b = 1$.

Comparison to Other Penalties in Statistics: Existing nonconvex alternatives to our proposed biconvex penalty in the form of the MCP, SCAD and Bridge penalties all come

with an extra hyperparameter controlling how close the penalty gets to approximating $||.||_0$ directly. One advantage of the biconvex penalty proposed here is that it does retains only the single global penalty strength parameter, which in practice means it is possible to simply try the model at all pertinent parameter values via a sequence of warm starts. However, it does require specification of a prior on λ. While this can be a limitation for some applications, we have found the standard Half-Cauchy to be sufficient across ours. Gong et al. (2013) developed a general iteration scheme for difference of convex problems. See Polson et al. (2015) for more on proximal methods in statistics.

2.3 Deploying the prox operator with VISTA

Deploying this proximal operator for optimization is as simple as plugging Equation 14 into the proximal descent procedure (Equation 6), at least in theory. In practice, success with any gradient descent method relies on choosing good step sizes, preconditioners, and acceleration. We discuss the details of our optimization procedure in Supplementary 1.

3 The Sparse Bayesian Adaptive Lasso

The purpose of this section is to develop a Bayesian Lasso procedure which provides full uncertainty quantification and penalty coefficient adaptation. We achieve this by deploying the VISTA procedure of the previous section on a novel nonsmooth Variational Bayesian methodology which we call the Sparse Bayesian Adaptive Lasso.

3.1 Varitional Inference with Nonsmooth Cost

We begin this section with a discussion of the impossibility of achieving geometric sparsity in standard VB cost functions before developing a nonsmooth penalty.

3.1.1 Background: Variational Bayesian Inference

Variational Bayes searches for a distribution over $\boldsymbol{\theta}$ that is 1) analytically tractable, and 2) sufficiently close to the posterior such that what each has to say about the quantities of interest are approximately the same. It does so by defining an optimization problem $\underset{Q_\theta \in \mathcal{Q}}{\operatorname{argmin}}\, d(Q_\theta, P_{\theta|\mathbf{y}})$ where $P_{\theta|\mathbf{y}}$ is our true posterior probability with density p and Q_θ is a candidate variational distribution with density q from a space of possible distributions $\mathcal{Q}$ parameterized by *variational parameters*, and d is some measure of dissimilarity between distributions. In this article, for each parameter θ, the variational distribution is defined by a location parameter η_θ and a scale parameter ν_θ. We will impose variational independence as is common: $q_{\boldsymbol{\eta},\boldsymbol{\nu}}(\boldsymbol{\theta}) = \prod_{m=1}^{M} q_{\eta_m,\nu_m}(\theta_m)$.

The most common d is the KL divergence from variational to posterior:

$$d(Q,P) = \mathrm{KL}(Q_\theta || P_{\theta|\mathbf{y}}) = \mathbb{E}_{\boldsymbol{\theta}\sim Q}[\log\left(\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{y})}\right)] = -\mathbb{E}_{\theta\sim Q}[\log \mathcal{L}(\mathbf{y}|\boldsymbol{\theta})] + \mathrm{KL}(Q_\theta||P_\theta)\,. \quad (16)$$

We see this $d(Q_\theta, P_{\theta|\mathbf{y}})$ may be viewed as the negative expected log likelihood (which we can view as a model misfit term) penalized by the KL divergence between the variational and prior distributions (which we can view as a model complexity term). Since Q was chosen to be simple, we can use a Monte Carlo sample to estimate the expected likelihood.

For the regression coefficients β_p, we specify Laplace distributions both for variational and prior distributions: $\beta \overset{Q}{\sim} \mathrm{L}(\eta_\beta, \nu_\beta)$ and $\beta \overset{P}{\sim} \mathrm{L}(0, \frac{1}{\lambda\tau})$, leading to the following penalty function on our variational parameters η_{β_p} and ν_{β_p}:

$$g_{\mathrm{KL}}(\eta_{\beta_p}, \nu_{\beta_p}, \lambda_p) = \mathrm{KL}(\mathrm{L}(\eta_{\beta_p}, \nu_{\beta_p})||\mathrm{L}(0, \frac{1}{\lambda\tau})) = \tau\lambda(\nu_{\beta_p} e^{-\frac{|\eta_{\beta_p}|}{\nu_{\beta_p}}} + |\eta_{\beta_p}|) - \log(\nu_{\beta_p}) - \log\lambda\,. \quad (17)$$

3.1.2 The Unbearable Smoothness of Variational Inference

One might be forgiven for thinking that Variational Bayes, a procedure built on optimization, would be the approach that allows us to combine the generality afforded by Bayesian

inference with the sparsity of penalized likelihood procedures. Unfortunately, the KL divergence, defined by an integral, will be continuously differentiable when any continuous prior distribution, such as the Laplace, is specified. As Fan and Li (2001) discuss in the context of penalized likelihood, a lack of nonsmoothness foils our hopes of sparse solutions.

The penalty given in Equation 17 is smooth despite the two absolute values because they cancel at zero as visualized in Figure 3 and demonstrated algebraically by Meyer (2021), who studied this function because of its similarity to the Huber loss function.

3.1.3 Variational Inference with a Nonsmooth Approximation

Owen (2007), in the context of penalized M estimation, advocates to reintroduce nonsmoothness by exchanging the quadratic and linear parts of the penalty. We propose to approximate $\nu_{\beta_p} e^{-\frac{|\eta_{\beta_p}|}{\nu_{\beta_p}}} + |\eta_{\beta_p}|$ by $\nu_{\beta_p} + |\eta_{\beta_p}|$ since it is a good approximation for small ν_{β_p}:

$$g_{\mathrm{NS}}(\eta_{\beta_p}, \nu_{\beta_p}, \lambda_p) := \tau\lambda_p(\nu_{\beta_p} + |\eta_{\beta_p}|) - \log(\nu_{\beta_p}) - \log\lambda_p - 1\,. \tag{18}$$

Figure 3 show that g_{NS} retains the qualitative properties of the KL penalty, and $g_{\mathrm{NS}}(\eta_{\beta_p}, \nu_{\beta_p}) - 1$ bounds $g_{\mathrm{KL}}(\eta_{\beta_p}, \nu_{\beta_p})$ from below. Notice that the term $\log \eta_{\beta_p}$ serves as a logarithmic barrier preventing $\eta_{\beta_p} \to 0$. The sparsity, therefore, is only on the posterior location parameter η_{β_p}, not the scale ν_{β_p}. Critically, this leads to qualitatively different sparsity than, for example, Spike-Slab or Horseshoe priors, which will shrink the entire posterior essentially to 0.

In the case of a linear regression with Gaussian error, the cost is closed form whenever the first two moments of $Q_{\boldsymbol{\beta}}$ and the inverse and log expectation of Q_{σ^2} are tractable (Supplementary 2). We cannot expect this to be the case in general, and must estimate the expected log likelihood term $-\mathbb{E}_{\boldsymbol{\theta}\sim Q_{\boldsymbol{\theta}}}[\log \mathcal{L}(\mathbf{y}|\boldsymbol{\theta})]$. Most common in "blackbox" VB (Ranganath et al., 2014) is to estimate this via a Monte Carlo sample at each iteration and to

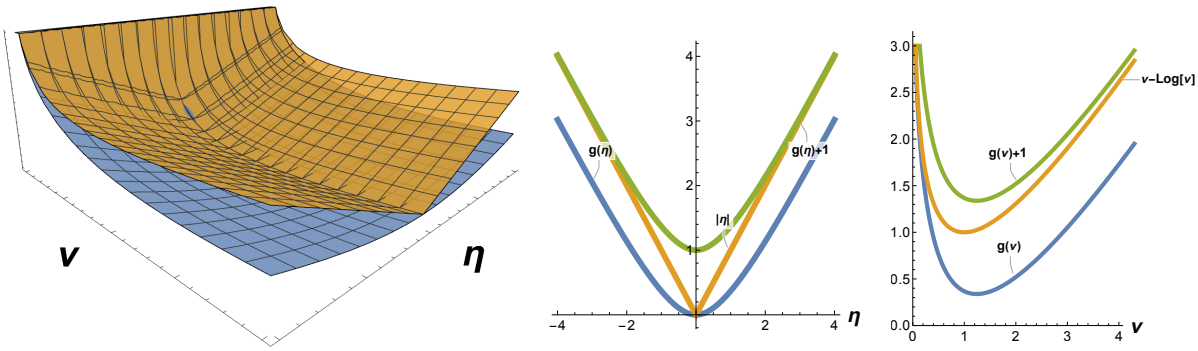

Figure 3: **Nonsmooth Approximation** *Left:* Bivariate KL penalty (blue) and nonsmooth approximation (orange). The marginal penalties for *Center:* η, g gives KL; *Right:* ν.

use stochastic optimization. We instead use the Sample Average Approximation (Robinson, 1996, SAA), which converts the optimization to a deterministic one by employing the same Monte Carlo draw throughout the optimization procedure, using antithetic sampling for variance reduction (Owen, 2013, Chapter 8.2). Collecting those parameters of interest which are not desired to be sparse (in our Case Study of Section 4 these include variance components, spatiotemporal correlations, and lag and zero inflation parameters) into the vector $\boldsymbol{\theta}$ with associated variational parameters η_θ, ν_θ, our cost is given as:

$$\min_{\boldsymbol{\eta}} \frac{1}{B} \sum_{\theta_b \sim Q_\theta, \beta_b \sim Q_\beta} \mathcal{L}(\theta_b, \beta_b) + \sum_{p=1}^{P} g_{\text{NS}}(\eta_{\beta_p}, \nu_{\beta_p}, \lambda_p) + \text{KL}(Q_{\boldsymbol{\alpha}} || P_{\boldsymbol{\alpha}}) \tag{19}$$

Note that to compute gradients we must differentiate through the sampling procedure; this is generally implemented in automatic differentiation packages like `tensorflow` and `pytorch`, though quality of the gradients vary. When the distribution is a location-scale family, this differentiation is particularly straightforward (Kingma and Welling, 2014). With the gradient of the smooth part computed, we can apply the VISTA algorithm to yield sparse learning with any likelihood. Since λ_p is of secondary interest, we estimate it via maximization rather than specifying a distribution for it.

3.2 Qualitative and Empirical Frequentist Properties

In this section, we examine the properties of this method in linear and generalized linear models. Together with our Laplace variational distribution on β_p, we place LogNormal priors and variational distributions over the error variance or overdispersion parameters σ^2. We begin by examining basic MAP in linear models before investigating consequences of the prior's nonconvexity using Subgradient and Proximal Gradient Descent, and conclude with a simulation study confirming the empirical coverage of Credible Intervals.

Working example $\mathbf{X} \in \mathbb{R}^{N \times P} = \mathbb{R}^{250 \times 10}$ with entries sampled from $i.i.d$ standard Gaussians, $\boldsymbol{\beta} \in \mathbb{R}^{50}$ with only 6 nonzero entries, of $[-2.5, -2, -1.5, 1.5, 2, 2.5]$ (similar β as Ročková and George (2018)), and $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with again iid standard normal ϵ_n. Figure 4, a shows a more readable problem with only $[-2, 2]$ nonzero in β and $N = 100$.

3.2.1 Effect of Prior Specification on Posterior Maxima in Linear Problems

Consider the generative model in the Gaussian case $\lambda \sim p_\lambda$; $\beta_p|\lambda_p, \tau \sim \mathrm{L}(0, \frac{1}{\tau\lambda_p})$; $y|\boldsymbol{\beta}, \sigma^2 \sim \mathrm{N}(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2)$, where, as in the Horseshoe, λ_p functions as a local scale parameter and τ as a global scale parameter. In this article, we will treat τ as known, varying it in applications in order to form a trajectory of parameter estimates interpolating between the model with all $\beta_p = 0$ and an unpenalized model. As we will see, increasing τ leads to greater coefficient sparsity.

We'll begin by developing the posterior density:

$$\max_{\boldsymbol{\beta} \in \mathbb{R}^P, \boldsymbol{\lambda} \geq \mathbf{0}} -\frac{||\mathbf{y} - \mathbf{X}\beta||_2^2}{2\sigma^2} + \sum_{p=1}^{P} \big[-\tau\lambda_p|\beta_p| + \log\lambda_p + \log\tau \big] + \sum_{p=1}^{P} \log p_\lambda(\lambda_p) \iff \tag{20}$$

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^P, \boldsymbol{\lambda} \geq \mathbf{0}} c(\boldsymbol{\beta}, \boldsymbol{\lambda}) := \frac{||\mathbf{y} - \mathbf{X}\beta||_2^2}{2\sigma^2} - \sum_{p=1}^{P} \big[\log\lambda_p + \log p_\lambda(\lambda_p) \big] - P\log\tau + \sum_{p=1}^{P} \tau\lambda_p|\beta_p| \tag{21}$$

Uniform $p_\lambda \propto 1$: Without a great idea of what prior we ought to put on λ, we might be

tempted to place a uniform prior on it. However, this leads to an unbounded density as $(\boldsymbol{\beta}, \boldsymbol{\lambda}) = (\mathbf{0}, \infty)$: the term $\frac{||\mathbf{y}-\mathbf{X}\boldsymbol{\beta}||_2^2}{2\sigma^2}$ becomes $\frac{||\mathbf{y}||_2^2}{2\sigma^2}$, while the $-\log \lambda_p$ terms are unbounded.

Undeterred, we performed MAP inference with a uniform prior on our working problem using the Adam optimizer (Kingma and Ba (2015)) initialized with $\boldsymbol{\beta} = \mathbf{0}$, $\boldsymbol{\lambda} = \mathbf{1}$ and subgradient descent, where it is seemingly able to perform debiased variable selection (see Figure 4a). The "zero $\boldsymbol{\beta}$ big $\boldsymbol{\lambda}$" strategy only surpasses the posterior density of local estimates on this particular example at $\lambda \approx 10^{200}$; since the $\log \lambda$ term is unbounded, we expect this behavior for any $\mathbf{X}, \mathbf{y}, \sigma^2$. Evidently, local optimizations may converge to a reasonable local minimum despite this. Though the biconvex penalty proposed here has its advantages over convex penalties, nonconvexity can be a treacherous property.

Power Inverse $p_\lambda \propto \frac{1}{\lambda^a}$: In this case, setting $\boldsymbol{\beta} = 0$ leaves us with a cost of $\sum_{p=1}^{P}(a-1)\log \lambda_p$. We see that for $a < 1$, we get similar behavior to the uniform case, where cost is unbounded below as $\lambda \to 0$. $a > 1$ leads to a positive coefficient for the $\log \lambda$ term, in turn leading to the opposite behavior, where cost is unbounded below as $\lambda \to 0$. When $a = 1$, the log terms cancel, and we are left with a cost of $\frac{||\mathbf{y}-\mathbf{X}\boldsymbol{\beta}||_2^2}{2\sigma^2} + \sum_{p=1}^{P} \lambda_p |\beta_p|$. Without any motivation to stay positive, the λ_p term will vanish, yielding the unpenalized problem.

We investigate the Half-Normal and Half-Cauchy cases under orthogonal designs in Supplementary 3.

3.2.2 Building Coefficient Trajectories by Varying τ

A very practically useful aspect of Lasso regression is the ability to efficiently form trajectories of parameter estimates while varying the penalty strength τ, warm-starting optimizations for efficiency. We also deploy VISTA estimating the SBL with Gaussian error structure and a Half-Cauchy prior, illustrated in Figure 4b, evaluate a trajectory for varying τ, and then compare to the Lasso on our working example. Though it successfully thresh-

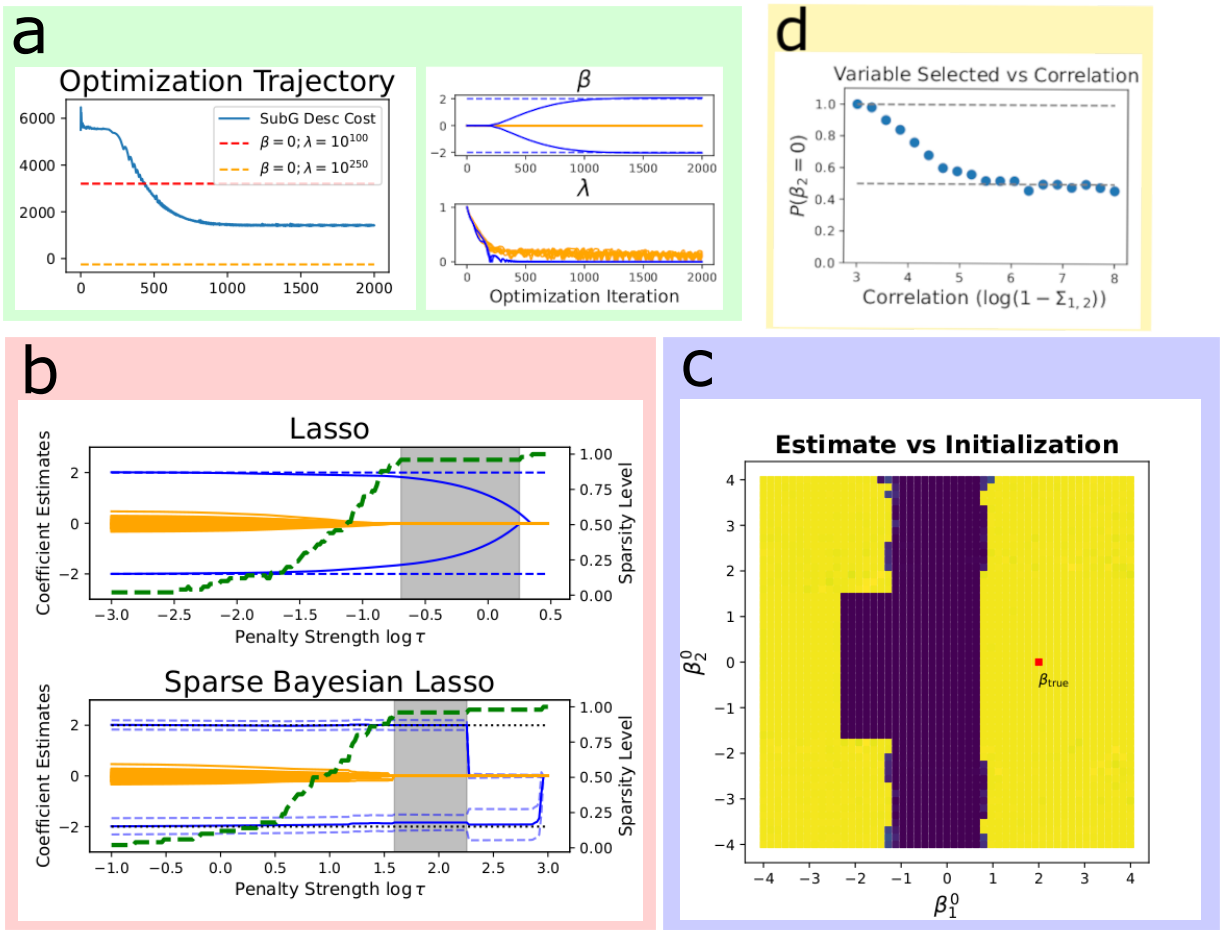

Figure 4: **Nonconvex Behavior.** *Top Left and Mid:* MAP inference with uniform prior leads to reasonable local optima, but likelihood is unbounded. *Lower and Mid Left:* Penalty τ varies: gray region is correct sparsity level, Lasso is biased within; SBL is debiased (like nonconvex penalties for likelihoods) and provides credible intervals. *Top Right:* Instability in RNG for selection of correlated parameters. *Bottom Right:* Cost vs Initialization, true parameter value in red.

olds the active parameters last, there is no regularization strength which allows Lasso to identify the correct model without bias. In the SBL, by contrast, the active parameters are near their true values for the whole range of penalty values until they are thresholded out, closer to the behavior of nonconvex penalties such as MCP and SCAD.

3.2.3 Nonconvexity and Instability

Since our penalty is nonconvex, the final parameters of an optimization are dependent on their initializations. Figure 4c shows the final costs associated with a grid of initializations on a 2D toy problem, where $\beta = [2, 0]$. The initial values are used on a grid of 51 points between [-4,4] and with τ set to 100. We perform MAP inference to focus on the nonconvexity in the coefficients (rather than on variational standard deviations). We see that in this case, the model lands on the true solution when β_1 is initialized at large values. For this simulation, and for the remainder of this article, λ_p were all initialized at 1 and β at 0. The classical Lasso is famously unstable insofar as it essentially randomly selects from groups of correlated parameters. We conduct a numerical experiment to confirm the inheritance of this property, using the same toy problem, but now with $\mathbf{X}$ data simulated from a correlated normal distribution with high correlation, and this time using the SBL. Figure 4d shows how, starting at a correlation of 0.999, as we approach a correlation of 1, the model begins changing from nearly always selecting β_1 as nonzero to doing so only 50% of the time.

3.2.4 Empirical Frequentist Properties

In this section, we examine the empirical performance of our proposed method on selected GLMs. We use the mean function described in our working example. We deploy the VISTA algorithm, implemented in `tensorflow_probability`, on for both MAP and VB inference

Likelihood	Method	beta-FNR	beta-FPR	beta-COVR	sig-COVR	TIME
bernoulli	VISTA-MAP	0.03	0.00	-	-	4.43
	VISTA-SBL	0.00	0.00	0.91	1.00	31.58
	Horseshoe	0.00	0.04	0.93	1.00	400.09
cauchy	SSLasso	0.88	0.00	-	-	0.02
	VISTA-MAP	0.30	0.00	-	-	6.43
	VISTA-SBL	0.00	0.03	0.94	0.97	40.86
	Horseshoe	0.00	0.00	0.94	0.87	458.67
nb	VISTA-MAP	0.00	0.02	-	-	6.31
	VISTA-SBL	0.00	0.00	0.95	1.00	28.77
	Horseshoe	0.00	0.00	0.94	1.00	962.86
normal	SSLasso	0.00	0.00	-	-	0.02
	VISTA-MAP	0.00	0.02	-	-	7.12
	VISTA-SBL	0.00	0.00	0.94	1.00	34.93
	Horseshoe	0.00	0.00	0.95	1.00	39.10
poisson	VISTA-MAP	0.00	0.00	-	-	18.01
	VISTA-SBL	0.00	0.00	0.80	1.00	46.77
	Horseshoe	0.00	0.00	0.98	1.00	394.34

Table 1: 95% Interval Coverage and Variable Selection performance on selected GLMs.

with Bernoulli, Negative Binomial, Normal, and Poisson likelihoods using the canonical link functions and with the Cauchy likelihood with an identity link. We tuned sparsity level τ by hand which is what we advocate doing in practice by examining the entire trajectory produced by varying τ Supplementary **??**. We use a Monte Carlo sample size of 40. As a comparator, we show a `JAGS` implementation of the Horseshoe prior, run for 1,000 iterations after adaptation on two chains, which should represent a lower bound on the amount of computation required for inference in practice. Additionally, we compare to the Spike-Slab Lasso using the R package `SSLASSO` when the link function is the identity, namely on the Gaussian and Cauchy cases. The Spike-Slab Lasso penalty could of course be combined with a non-Gaussian likelihood, but solving the marginal problem for coordinate descent will no longer in general be in closed form and is not implemented in `SSLASSO`. In contrast, the proximal operator is the same for any differentiable likelihood.

As Table 1 shows, SBL is able to maintain high interval coverage using generally about

an order of magnitude the elapsed real time of the Horseshoe prior. On the Negative Binomial likelihood, which is the likelihood we use in the case study of Section 4.3, we observe about a 30× speedup. The SSLasso is fast but not yet applicable to non-Gaussian likelihoods (estimation with Cauchy data goes poorly), and not able to provide UQ. The performance gap is smaller on the Normal likelihood where the Gibbs sampler enjoys conjugacy.

4 Modeling Migration in Iraq

We begin this section with an overview of the classical Gravity Model. We then pivot to discussing the data set we collected before variable selection. Finally, we present the results of our analysis.

4.1 ISIL and the 2013-2017 Iraqi Civil War

In late 2013, tensions between Sunni and Shia Muslims in the Anbar province of Iraq boiled over into violence between the Shia government and its allies against the Islamic State. The violence led to significant internal displacement which is the subject of our case study. In the absence of traditional migration variables, we consider the use of social media for capturing indirect indicators of migrations. Specifically, we focus on novel conversation buzz and insecurity predictors based on Twitter data.

4.1.1 Data Creation and Collection

Social Media Indicators We collected Arabic language tweets between March 2016 and June 2017 from Twitter's Streaming Application Program Interface (API). Buzz variables were formed by monitoring counts of keywords selected by Arabic language experts re-

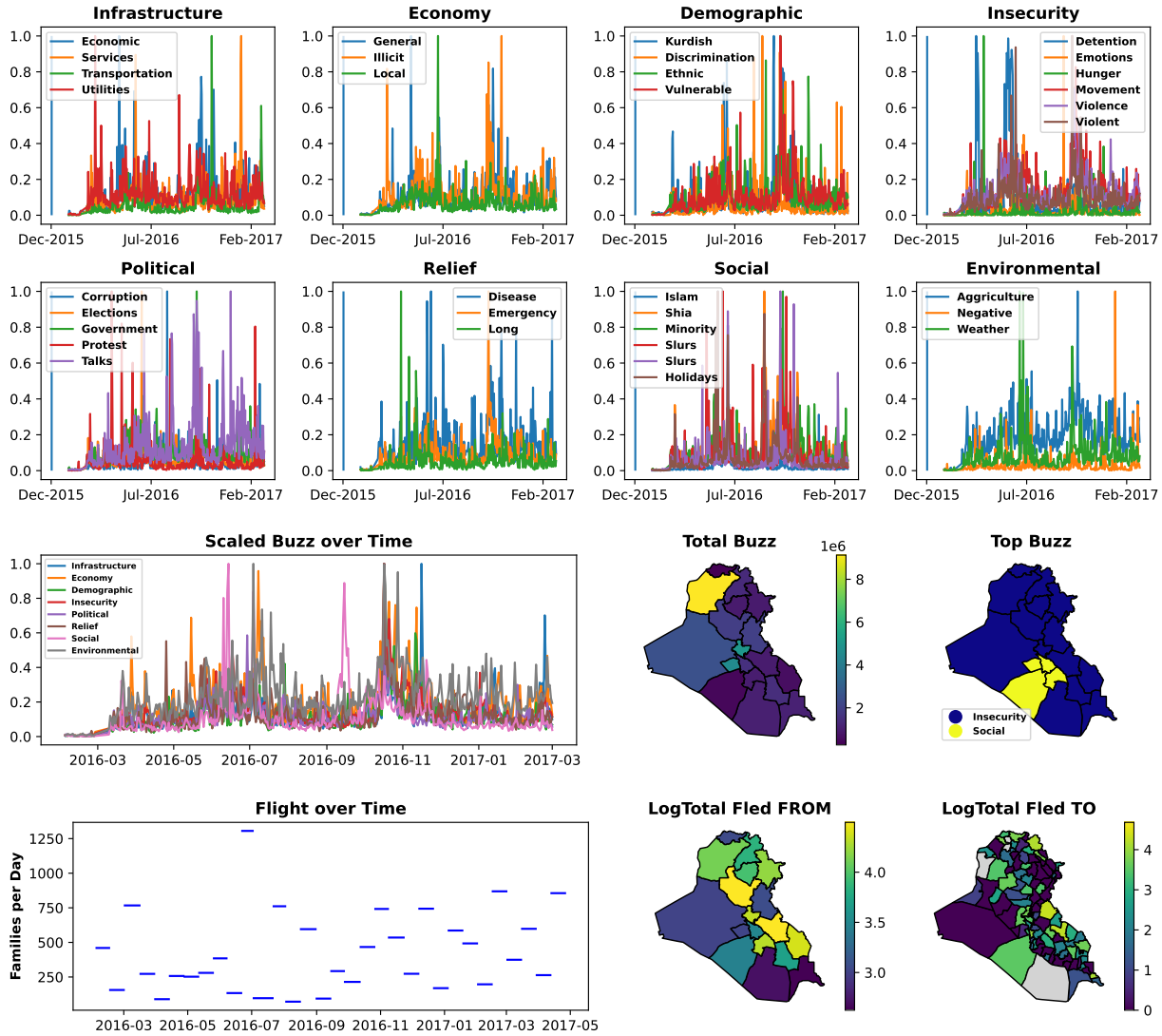

Figure 5: **Iraq Dataset:** *Top Two Rows:* Buzz Variables broken up by category and normalized to have max value 1. *Bottom Left Quadrant:* The cumulative buzz by each category is shown above the UN IOM data giving the families fleeing during each period. *Bottom Right Quadrant:* Spatial totals for Buzz (upper row) and flow (lower row); gray districts not observed.

lated to various migration-relevant topics, such as ethnic angst and discussion about the economy. Sentiment associated with these topics was also captured. Variables such as Social buzz for different demographic groups, (e.g. `Social-shia_buzz`), Insecurity, (e.g. `Insecurity-emotions`), Environmental (e.g. `Environmental-weather`) are examples of the 68 variables identified.

Movement Data We use the flow data provided by the International Organization for Migration's Iraq Displacement Tracking Matrix, which gives stocks of displaced families in one of 97 Districts, as well as which of the 18 Governorates they originated from. We differenced these data over time to estimate flows. The counts are aggregated over variable length periods, with the shortest period consisting of 11 days and longest of 31 days. There are 48 total time periods in our study, leading to a total of 57,024 observations. 96.4% of response values are zero, with a mean of 6.71 families displaced for all time across origin-destination pairs.

4.1.2 New Challenges for the Gravity Model

Gravity models of migration (Poprawe, 2015; Karemera et al., 2000; Ramos and Suriñach, 2017) measure the exchange between the origin o and destination d for a given time period t, denoted as $y_{t,o,d}$; in our case study $y_{t,o,d}$ is measured in number of families. The gravity model specifies a generalized linear model relating $\mathbf{y} \in \mathbb{R}^{N_o N_d T}$ to predictors $\mathbf{X} \in \mathbb{R}^{N_o N_d T \times P}$ as well as *resistance to exchange* terms ω_t, ω_o, ω_d which represent the effects associated with each time periods, origin and destination. Poisson gravity models have been fit using MCMC for hospital patient flows (Congdon, 2000) and trade (Ranjan and Tobias, 2007). Chen et al. (2018) perform explicit Bayesian Model Selection on a gravity model of trade via Normal approximations to their likelihood. The fine timescale of our data introduces temporal misalignment between cause and effect: if an event which causes migration is

recorded by our Twitter Buzz variables on a given day, we do not expect migrants to register at a camp the same day. At the yearly scale which gravity models are typically deployed, this is not a concern, but we need to develop a mechanism for estimating lags in this specific case study.

4.2 Spatiotemporal ZINB Gravity Model with Gaussian Lags

In this section, we develop the following hierarchical Bayesian model to relate the organic variables to the observed migration:

$$\lambda_p \overset{\text{iid}}{\sim} \text{C}(0,1)^+ \text{ for } p \in \{1,\dots,P\} \tag{22}$$

$$\beta_p|\lambda_p \overset{\text{indep}}{\sim} \text{L}(0, \frac{1}{\tau\lambda_p})^+ \text{ for } p \in \{1,\dots,P\} \tag{23}$$

$$\phi_{\mathcal{I}} \sim \text{logitN}(a_\phi, b_\phi) \text{ for } \mathcal{I} \in \{\mathcal{T},\mathcal{O},\mathcal{D}\} \tag{24}$$

$$\rho^2_{\mathcal{I}} \sim \text{logN}(a_\rho, b_\rho) \text{ for } \mathcal{I} \in \{\mathcal{T},\mathcal{O},\mathcal{D}\} \tag{25}$$

$$\boldsymbol{\omega}_{\mathcal{I}}|\phi_{\mathcal{I}}, \rho^2_{\mathcal{I}} \sim \text{N}\big(\mathbf{0}, \rho^2_{\mathcal{I}}\Sigma_{\mathcal{I}}(\phi_{\mathcal{I}})\big) \text{ for } \mathcal{I} \in \{\mathcal{T},\mathcal{O},\mathcal{D}\} \tag{26}$$

$$\pi \sim \text{logitN}(a_\pi, b_\pi) \tag{27}$$

$$y_{t,o,d} \sim \begin{cases} \delta_0 & \text{w.p. } \pi \\ \text{N.B.}(g(\mathbf{z}_{t,o,d}(\mu_l,\sigma_l)^\top\beta + \omega_t + \omega_o + \omega_d), \frac{l_t}{\sigma^2}) & \text{o.w.} \end{cases} \tag{28}$$

where $\phi_{\mathcal{T}} \in [0,1]$ is the correlation between successive temporal fixed effects $\boldsymbol{\omega}_t$ giving the AR1 correlation matrix $\Sigma_t(\phi_t)$, while ϕ_o, ϕ_d are the correlation parameters for the origin and destination spatial effects $\boldsymbol{\omega}_{\mathcal{O}}, \boldsymbol{\omega}_{\mathcal{D}}$, respectively, which are scaled to lie between the inverse maximum and minimum eigenvalues of the adjacency matrix (De Oliveira, 2012). These matrices are scaled by variance terms $\rho_{\mathcal{I}}$ which are endowed with LogNormal priors. Given all these quantities, the number of families moving between two regions on a given

day is given by a negative binomial distribution with success probability given by the canonical link transformation g of our linear predictions. The positive "number of failures" parameter controlling overdispersion is estimated by $\frac{1}{\sigma^2}$ and is scaled by the duration of the time period l_t. Due to the preponderance of zeros in our model, we include the possibility of zero inflation by specifying a zero inflation parameter π with logitNormal prior. We account for unknown and variable lag between cause and effect through a "flattened" Gaussian filter with parameter μ_l determining how far back to shift the time series and the parameter σ_l determining how far outside the bounds of the period length to integrate (Supplementary 4.2). We use a neighborhood structure that has weight 1 if two regions touch and 0 otherwise.

4.2.1 Variational Distribution and Inference

Parameters $\boldsymbol{\lambda}$, $\boldsymbol{\beta}$ and σ^2 are estimated as in Section 3.2. We place logitNormal variational distributions on $\phi_t, \phi_o, \phi_d, \pi$ and logNormal variationals on $\rho_t, \rho_o, \rho_d, \sigma^2$. The KL divergence between prior and variational in all of these cases is given by the KL divergence of the underlying normal distributions and thus is available in closed form. The ω's are given a Normal variational distribution. While the KL divergence cannot completely be determined in closed form due to the spatiotemporal correlation parameters $\boldsymbol{\phi}$, the ω and ρ integrals may be computed analytically, and this expression may be evaluated for Monte Carlo samples of ϕ without any matrix decompositions beyond an initial decomposition of the neighborhood structure matrix (Supplementary 4.2). As is done in the context of Gibbs sampling for CAR models, we "center on the fly" (Ferreira, 2019) by constraining the variational mean of $\boldsymbol{\omega}_{\mathcal{I}}$ to have mean 0 after each proximal gradient iteration: $\boldsymbol{\eta}_{\omega_{\mathcal{I}}} \leftarrow \boldsymbol{\eta}_{\omega_{\mathcal{I}}} - \text{mean}(\boldsymbol{\eta}_{\omega_{\mathcal{I}}})\mathbf{1}$.

4.3 Results

In this section, we apply our methodology to our case study. We begin by setting the regularization coefficient $\tau = 250$, where the model chooses to include only the destination-level variable `insecurity-emotions-buz`. We then decreased this coefficient along a logarithmically spaced grid to $\tau = 10$ with 100 points in between, plotting the trajectory of the variational distributions in Figure 6. Figure 6 shows the trajectory borne out by the regression coefficients, random effect variances and lag and zero inflation parameters. Begin with the Left-Top three panels, which show the variational distributions of the regression coefficients. Qualitatively, we notice that variances of zeroed out coefficients decrease smoothly as the regularization strength increases, reflecting the increased prior distribution concentration. Nonzero coefficients, however, tend to have relatively stable values before experiencing jumps as new parameters enter. The first variable to enter the model is the `insecurity-emotions-buzz` variable associated with the destination. The next variable to enter is `road-dist`, the dyadic distance between origin and destination. It is not surprising that this variable enters early: it is what gives the gravity model its name. But it is somewhat unusual for it to have a positive coefficient: this is because of the unique geographical characteristic of the Iraq conflict (see Section 4.4 for more discussion). The next variables to enter are again at the Destination level: `relief-aid` with a positive coefficient, which could indicate either that people want to travel to places with good relief aid, or simply that places people are displaced to are where the aid workers locate themselves, as well as `environment-negative` sentiment with a negative coefficient, indicating that the environment related conversation around some locations has a negative tone and are highly correlated to areas where migrants are not moving. The `environment-negative` sentiment at the Origin level is the next variable to enter, this time with the opposite

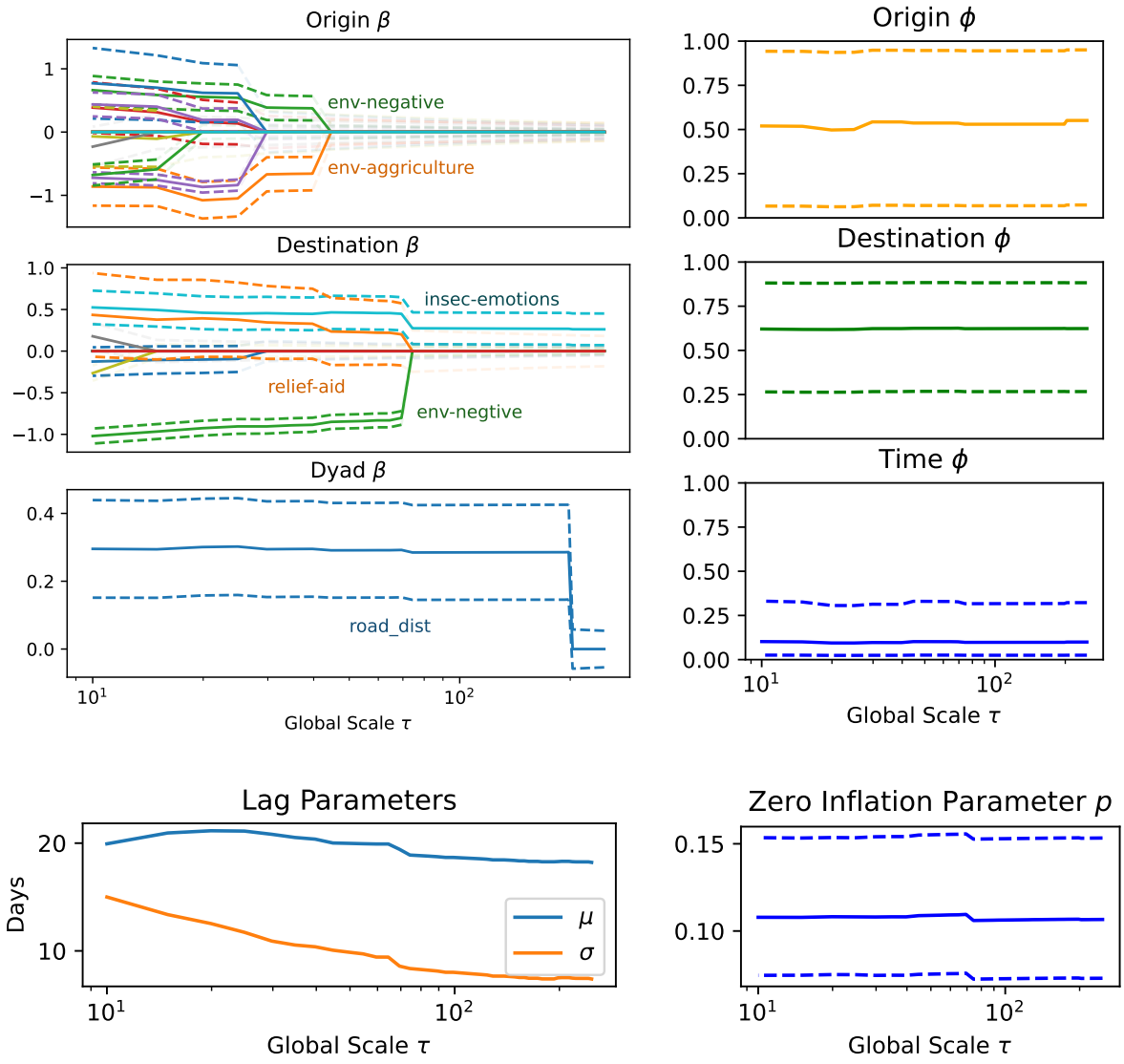

Figure 6: **Parameter Trajectories vs τ.** *Top Left:* Regression Coefficient Variational Trajectory; dotted lines give 95% credible intervals; intervals are faded when the posterior mean is 0. *Top Right:* Spatiotemporal Correlations fairly constant. *Bottom Left:* Gaussian convolution parameters vary with fixed effects. *Bottom Left:* Zero inflation parameter is constant.

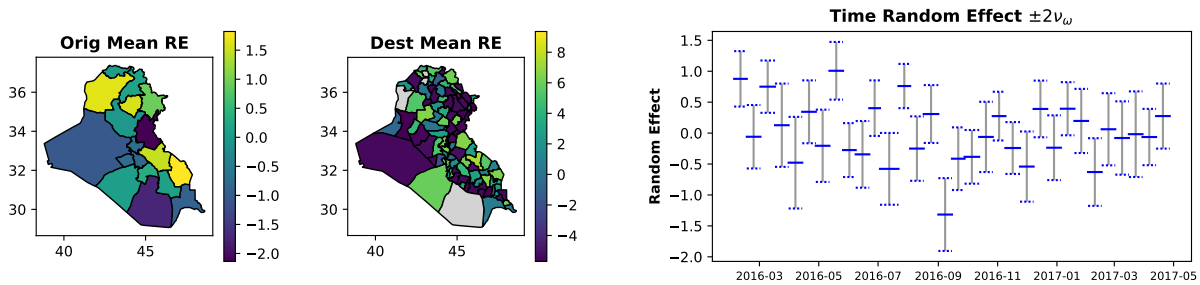

Figure 7: **Spatiotemporal Random Effects:** At the largest τ value.

coefficient, simultaneously with the `environment-agriculture` buzz at the Origin with a negative coefficient, perhaps indicating that general buzz or discussion is a correlate to environmental conditions that reduce migration while negative sentiment about environmental conditions in other locations is a correlate to an environmental push factor for migration.

We now discuss the random effects, which are relatively stable as τ changes. The lag-aggregation parameters are shown in the bottom left of Figure 6. We see that as the number of variables in the model increases, so too does the lookback window size σ_l, from about 9 days to about 15. The number of days to look back travels from around 18 to around 22. These results indicate that we can choose between a model with sharper predictions from fewer variables, or smoother predictions incorporating more variables and time. The spatiotemporal correlation parameters' variational distributions are shown in the top 3 right panels of Figure 6. The origin parameter is not well pinned down by our only 18 governorates, the Destination is also not strongly identified though it does rule out substantial negative spatial correlation, and the temporal correlation seems minimal. Finally, the bottom right panel indicates that the zero-inflation parameter is also fairly constant with respect to τ, and somewhere between 8 and 15 percent. Recall that the data were over 95% zeros; most of these are being accounted for by the mean or correlation structure rather than the zero inflation.

One of the biggest and most positive origin random effects belongs to Ninewa Gover-

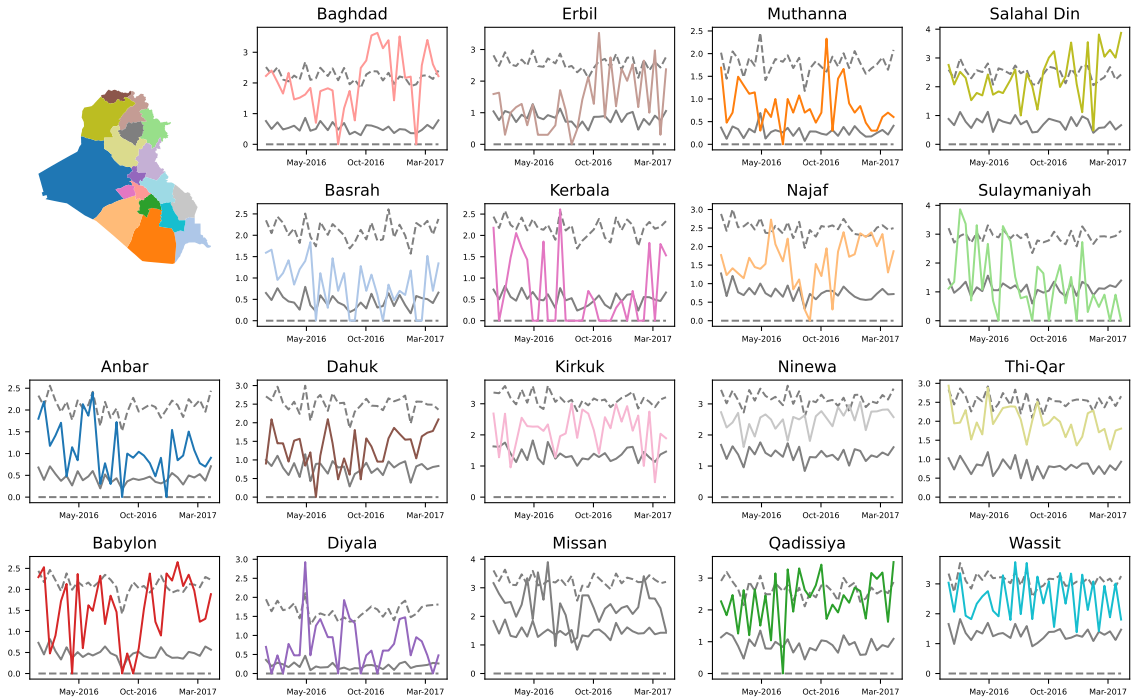

Figure 8: **Prediction by Origin Governorate** 95% variational predictive intervals appear to have uncertainty calibrated to the observed flows.

norate. Ninewa's capital is Mosul, a city which was the subject of fierce urban combat against ISIS, so we expect a positive random effect indicating increased migration from there. Maysan and Kirkuk governorates also large positive effects. The most negative effects belong to the Diyala Governorate, a Kurdish province in the East, and the Al-Muthanna Governorate, a southern province.

For destinations, on the other hand, the Baghdad area has some of the highest effects, with its constituent Al-Sadr and Kut districts taking places one and two and fairly well separated from the rest. On the lower end we see a different story: many districts, most of which observed no inflow of migration, have random effects which have settled into a value of about -5. Technically, these random effects are thus separated and not well defined by the likelihood; this is a known difficulty in gravity modeling (and count data in

general). Our spatiotemporal prior prevents them from wandering off to negative infinity, but cannot divine exactly what they should be. Nevertheless, the model assigns more variance to destination effects than to origin or temporal effects.

The time period with the largest random effect was from 2016-05-12 to 2016-05-26. According to a timeline from the Wilson Center (Center, 2019), this period saw a series of ISIS bomb attacks which left over 100 dead in Baghdad (May 12) as well as combat between Iraqi and ISIS forces. The period with the lowest random effect was 2016-09-01 to 2016-09-15, during which ISIS was more active abroad than in Iraq, also according to the Wilson Center's timeline.

5 Discussion and Future Work

We began this article by examining the density of the Bayesian Lasso when λ_p is given a hyperprior and treated as a variable to optimize. This lead to a biconvex proximal operator that enjoys singlevaluedness for sufficiently small step sizes. We then showed how to plug this proximal operator into the ISTA algorithm, producing the VISTA algorithm which allows for minimal bias even in the presence of high sparsity. Next, we examined the behavior of the MAP estimate in simple linear cases before developing a nonsmooth Variational Bayesian Lasso and deploying the VISTA algorithm on GLMs, where we found that our variational procedure had good frequentist coverage at less computational cost than simulation. Finally, we deployed the VISTA algorithm and Sparse Bayesian Lasso on a sophisticated hierarchical model and were able to generate trajectories of variational distributions of all model parameters, varying the penalty strength.

We look forward to deploying the VISTA algorithm on other applications such as the inverse and imaging problems where ISTA made its name. And we are even more excited

to see what other possibilities viewing penalty coefficients as optimization quantities can open up. In particular, Group Lasso (Yuan and Lin, 2006) and Fused Lasso (Tibshirani et al., 2005), among other Lasso analogs, enjoy fast inference via proximal gradient methods as well; perhaps the variable-penalty versions of these proximal operators are available in closed form. Variable-coefficient low-rank penalties for matrices (e.g. Koltchinskii et al. (2011)) and their associated proximal operators are also of future interest. Recently, Quaini and Trojani (2022) developed basic inferential properties of estimators defined via a proximal operator; it would be interesting to see which of these properties apply to the biconvex proximal operator we have developed here. We are also curious about different ways to motivate nonsmoothness in VB, perhaps via different divergences or discontinuous priors.

Acknowledgements

The authors would like to thank Stephen Becker, Warren Hare, Nick Polson, Mahlet Tadesse, Stefan Wild and Samy Wu Fung for valuable conversations and input.

References

Babacan, S. D., S. Nakajima, and M. N. Do (2014). Bayesian group-sparse modeling and variational inference. *IEEE transactions on signal processing 62*(11), 2906–2921.

Beck, A. and M. Teboulle (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences 2*(1), 183–202.

Bhadra, A., J. Datta, N. G. Polson, and B. Willard (2019). Lasso meets horseshoe: A survey. *Statistical Science 34*(3), 405–427.

Bhattacharya, A., D. Pati, N. S. Pillai, and D. B. Dunson (2012). Bayesian shrinkage. *arXiv preprint arXiv:1212.6088*.

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association 112*(518), 859–877.

Bühlmann, P. and L. Meier (2008). Discussion: One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics 36*(4), 1534 – 1541.

Candes, E. J., M. B. Wakin, and S. P. Boyd (2008). Enhancing sparsity by reweighted l 1 minimization. *Journal of Fourier analysis and applications 14*(5), 877–905.

Carvalho, C. M., N. G. Polson, and J. G. Scott (2010). The horseshoe estimator for sparse signals. *Biometrika 97*(2), 465–480.

Center, W. (2019). Timeline: the rise, spread, and fall of the islamic state.

Chen, H., A. Mirestean, and C. G. Tsangarides (2018). Bayesian model averaging for dynamic panels with an application to a trade gravity model. *Econometric Reviews 37*(7), 777–805.

Congdon, P. (2000). A bayesian approach to prediction using the gravity model, with an application to patient flow modeling. *Geographical analysis 32*(3), 205–224.

Daubechies, I., M. Defrise, and C. De Mol (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics 57*(11), 1413–1457.

De Oliveira, V. (2012). Bayesian analysis of conditional autoregressive models. *Annals of the Institute of Statistical Mathematics 64*(1), 107–133.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics 32*(2), 407 – 499.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Ferreira, M. A. (2019). The limiting distribution of the gibbs sampler for the intrinsic conditional autoregressive model. *Brazilian Journal of Probability and Statistics 33*(4), 734–744.

Gong, P., C. Zhang, Z. Lu, J. Huang, and J. Ye (2013). A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *international conference on machine learning*, pp. 37–45. PMLR.

Hare, W. and C. Sagastizábal (2009). Computing proximal points of nonconvex functions. *Mathematical Programming 116*(1), 221–258.

Hunter, D. R. and R. Li (2005). Variable selection using MM algorithms. *The Annals of Statistics 33*(4), 1617 – 1642.

Kang, J. and J. Guo (2009). Self-adaptive lasso and its bayesian estimation. Technical report, Working Paper.

Karemera, D., V. I. Oguledo, and B. Davis (2000). A gravity model analysis of international migration to north america. *Applied Economics 32*(13), 1745–1755.

Kawano, S., I. Hoshina, K. Shimamura, and S. Konishi (2015). Predictive model selection criteria for bayesian lasso regression. *Journal of the Japanese Society of Computational Statistics 28*(1), 67–82.

Kingma, D. P. and J. Ba (2015). Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Kingma, D. P. and M. Welling (2014). Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Koltchinskii, V., K. Lounici, and A. B. Tsybakov (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics 39*(5), 2302–2329.

Leng, C., M.-N. Tran, and D. Nott (2014). Bayesian adaptive lasso. *Annals of the Institute of Statistical Mathematics 66*(2), 221–244.

Makalic, E. and D. F. Schmidt (2015). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters 23*(1), 179–182.

Mallick, H. and N. Yi (2014). A new Bayesian lasso. *Statistics and its interface 7*(4), 571.

Marjanovic, G. and V. Solo (2013). On exact l q denoising. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6068–6072. IEEE.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis 52*(1), 374–393.

Meyer, G. P. (2021, jun). An alternative probabilistic interpretation of the huber loss. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA, pp. 5257–5265. IEEE Computer Society.

Mitchell, T. J. and J. J. Beauchamp (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association 83*(404), 1023–1032.

Owen, A. B. (2007). A robust hybrid of lasso and ridge regression. *Contemporary Mathematics 443*(7), 59–72.

Owen, A. B. (2013). *Monte Carlo theory, methods and examples.*

Parikh, N., S. Boyd, et al. (2014). Proximal algorithms. *Foundations and Trends in Optimization 1*(3), 127–239.

Park, T. and G. Casella (2008). The Bayesian lasso. *Journal of the American Statistical Association 103*(482), 681–686.

Polson, N. G., J. G. Scott, and B. T. Willard (2015). Proximal algorithms in statistics and machine learning. *Statistical Science 30*(4), 559–581.

Poprawe, M. (2015, Jun). On the relationship between corruption and migration: empirical evidence from a gravity model of migration. *Public Choice 163*(3), 337–354.

Quaini, A. and F. Trojani (2022). Proximal estimation and inference. *arXiv preprint arXiv:2205.13469*.

Ramos, R. and J. Suriñach (2017). A gravity model of migration between the ENC and the EU. *Tijdschrift voor Economische en Sociale Geografie 108*(1), 21–35.

Ranganath, R., S. Gerrish, and D. Blei (2014, 22–25 Apr). Black Box Variational Inference. In S. Kaski and J. Corander (Eds.), *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, Volume 33 of *Proceedings of Machine Learning Research*, Reykjavik, Iceland, pp. 814–822. PMLR.

Ranjan, P. and J. L. Tobias (2007). Bayesian inference for the gravity model. *Journal of Applied Econometrics 22*(4), 817–838.

Robinson, S. M. (1996). Analysis of sample-path optimization. *Mathematics of Operations Research 21*(3), 513–528.

Ročková, V. and E. I. George (2018). The spike-and-slab lasso. *Journal of the American Statistical Association 113*(521), 431–444.

She, Y. (2009). Thresholding-based iterative selection procedures for model selection and shrinkage. *Electronic Journal of statistics 3*, 384–415.

Taylor, H. L., S. C. Banks, and J. F. McCoy (1979). Deconvolution with the ℓ_1 norm. *Geophysics 44*(1), 39–52.

Terenin, A., S. Dong, and D. Draper (2019). GPU-accelerated Gibbs sampling: a case study of the horseshoe probit model. *Statistics and Computing 29*(2), 301–310.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267–288.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(1), 91–108.

Tung, D. T., M.-N. Tran, and T. M. Cuong (2019). Bayesian adaptive lasso with variational bayes for variable selection in high-dimensional generalized linear mixed models. *Communications in Statistics-Simulation and Computation 48*(2), 530–543.

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

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics 38*(2), 894 – 942.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association 101*(476), 1418–1429.

Zou, H. and R. Li (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics 36*(4), 1509 – 1533.

SUPPLEMENTARY MATERIAL

Python Package: PyPI-package `sbl` containing code to perform the the methods outlined in this article.

Supplementary Discussion and Results PDF Document containing additional numerical results, mathematical results, and background.