OPT2022: 14th Annual Workshop on Optimization for Machine Learning

A Variable-Coefficient Nuclear Norm Penalty for Low Rank Inference

Nathan Wycoff NATHAN.WYCOFF@GEORGETOWN.EDU
The McCourt School's Massive Data Institute, Georgetown University

Ali Arab ALI.ARAB@GEORGETOWN.EDU
Department of Mathematics and Statistics, Georgetown University

Lisa O. Singh LISA.SINGH@GEORGETOWN.EDU
The McCourt School's Massive Data Institute and Department of Computer Science, Georgetown University

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 $P = \min(M, N)$. Local solutions to this class of problems can produced efficiently for smooth f using proximal gradient methods [16]. Most prominently, Cai et al. [2] and Mazumder et al. [14] 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 estimated via optimization, allowing for a better surrogate loss function which more closely mimics the behavior of the rank function while maintaining a continuous optimization problem. The remainder of this article is organized as follows: Section 2 provides an overview of pertinent sparsity and low-rank inducing penalties. In Section 3, we describe the novel penalty and an optimization algorithm.

In Section 4, we deploy our penalty on noisy matrix completion problem based on natural images. Finally, Section 5 concludes and overviews future research.

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:

for $t \in \{1, \ldots, T\}$ **do**
  $\mathbf{x}_0^{t+1} \leftarrow \mathbf{x}^t - \eta \nabla f(\mathbf{x}^t)$
  $\mathbf{x}^{t+1} \leftarrow \mathrm{STO}_{\eta\tau}(\mathbf{x}_0^{t+1})$,
end

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

for $t \in \{1, \ldots, T\}$ **do**
  $\mathbf{A}_0^{t+1} \leftarrow \mathbf{A}^t - \eta \nabla f(\mathbf{A}^t)$
  $\mathbf{U}^{t+1}, \boldsymbol{\sigma}^{t+1}, \mathbf{V}^{t+1,\top} \leftarrow \mathrm{svd}(\mathbf{A}_0^{t+1})$
  $\mathbf{A}^{t+1} \leftarrow \mathbf{U}^{t+1} \mathrm{diag}[\mathrm{STO}_{\eta\tau}(\boldsymbol{\sigma}^{t+1})] \mathbf{V}^{t+1,\top}$
end

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 [21] is now ubiquitous in many domains of computational science, and is called *Lasso regression* [22] 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 [25, MCP], the Smoothly Clipped Absolute Deviation [7, SCAD] and bridge/ℓ_q [10] penalties $p(\mathbf{x}) = \lambda \sum_{p=1}^P |x_p|^q$, $q \in (0, 1)$. Yao et al. [24] and Phan and Nguyen [18] developed a general optimization framework that can accommodate many nonconvex penalties. Marjanovic and Solo [13] 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. Mohan and Fazel [15], on the other hand, develop an iterative reweighting procedure for the bridge norm, defining a weighted version of the nuclear norm: $\|\mathbf{A}\|_*^{\boldsymbol{\lambda}} = \sum_{p=1}^P \lambda_p \sigma_p(\mathbf{A})$. Lu et al. [12] 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 [17] 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 [20] 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 [23], 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 $\tau\eta^2 \geq 1$ (with x^* unchanged). In general the constraint is $\tau\eta_x\eta_\lambda \geq 1$ if these have different step sizes. 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, conducting MAP inference with this prior structure. The optimization problem thus becomes:

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

where $c(\boldsymbol{\lambda}) = \log\big[\int_{\mathbf{A}\in\mathbb{R}^{M\times N}} e^{-\|\mathbf{A}\|_*}\big]$ is the normalizing constant associated with this distribution. Unfortunately, we have not been able to locate a discussion of this distribution in the academic literature, and, for the purposes of our numerical experiments, simply plug in the normalization constant of the scalar Laplace $c(\boldsymbol{\lambda}) = \sum_p \log\lambda_p$. We intend to develop the normalizing expression in future work. 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 $\mathbf{A}$:

for $t \in \{1, \ldots, T\}$ **do**
  $\mathbf{A}_0^{t+1} \leftarrow \mathbf{A}^t - \eta\nabla_{\mathbf{A}} f(\mathbf{A}^t)$
  $\boldsymbol{\lambda}_0^{t+1} \leftarrow \boldsymbol{\lambda}^t - \eta\nabla_{\boldsymbol{\lambda}}\big[-\sum_{p=1}^{P}\log P_\lambda(\boldsymbol{\lambda}^t) - c(\boldsymbol{\lambda})\big]$
  $\mathbf{U}^{t+1}, \boldsymbol{\sigma}_0^{t+1}, \mathbf{V}^{t+1,\top} \leftarrow \mathrm{svd}(\mathbf{A}_0^{t+1})$
  $\boldsymbol{\lambda}^{t+1}, \boldsymbol{\sigma}^{t+1} \leftarrow \mathrm{VSTO}_{\tau\eta}(\boldsymbol{\lambda}_0^{t+1}, \boldsymbol{\sigma}_0^{t+1})$
  $\mathbf{A}^{t+1} \leftarrow \mathbf{U}^{t+1}\mathrm{diag}(\boldsymbol{\sigma}^{t+1})\mathbf{V}^{t+1,\top}$
end

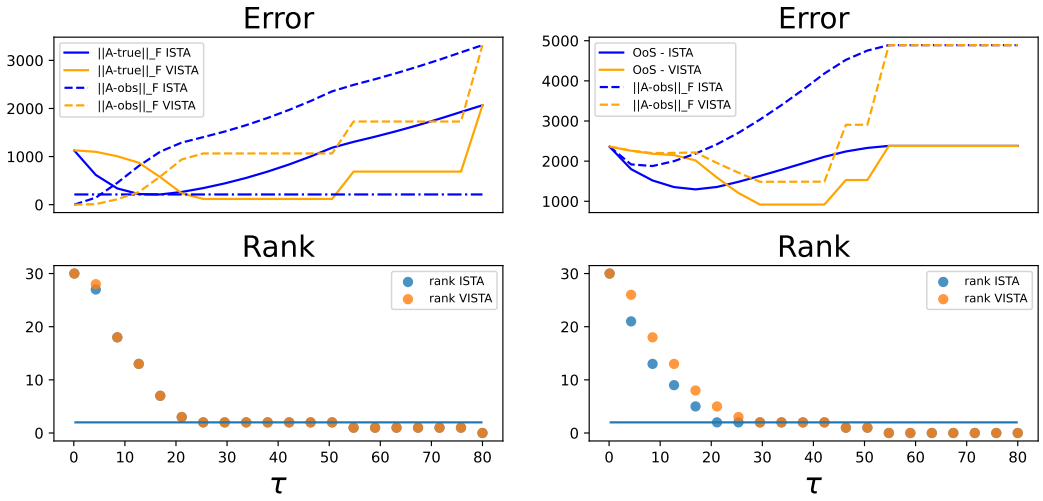

Figure 1: **Synthetic Illustration on Individual Matrices:** $\mathbf{A} \in \mathbb{R}^{30\times 40}$ with standard normal noise. *Left:* Denoising. *Right:* Noisy Matrix Completion. Dotted lines give reconstruction error, solid lines give prediction error.

Unlike the previously proposed nonconvex penalties enumerated by [12], 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 [23].

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.

A second difference is that the weighted nuclear norm is nonconvex [11], in contrast to the weighted ℓ_1 norm, which retains the convexity of the unweighted ℓ_1 norm. In either case, the problem with $\boldsymbol{\lambda}$ considered as a decision variable is nonconvex, but it in the vector case it is at least biconvex. In future work, we will investigate the implications of this additional nonconvexity in the matrix case.

4. Empirical Evaluation

In this section, we first qualitatively compare the proposed VISTA approach with the classical ISTA on two random matrices 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 com-

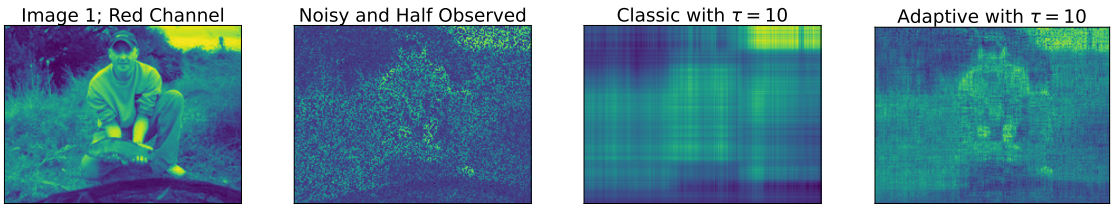

Figure 2: **Example Case:** *Left to Right:* Original Image, Noisy partial image presented to algorithms, ISTA solution, VISTA solution.

pletion and we are able to use images small enough to easily compute explicit singular values decompositions.

4.1. Experimental Design

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:

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

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.

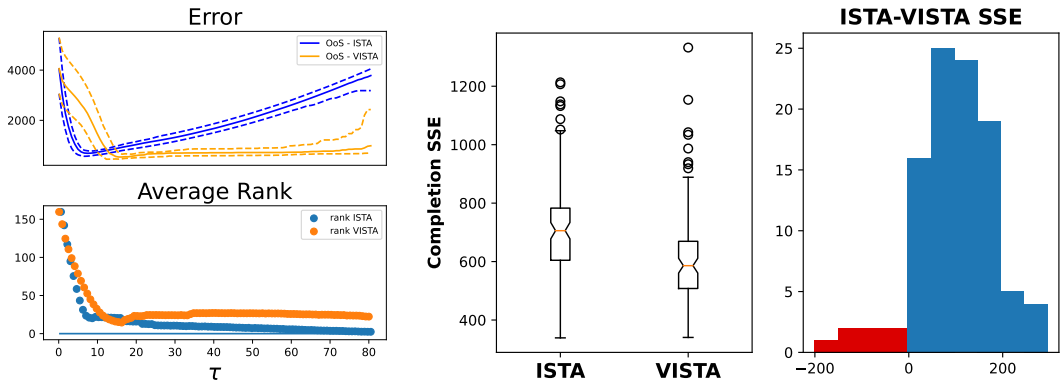

Figure 3: **Imagenet Test Problems:** *Left:* The plot gives the 2.5, 50, and 97.5 percentiles of sum squared prediction error for both algorithms for a range of τ. *Center:* The distribution of out of sample SSE for VISTA is below that of ISTA. *Right:* A histogram of the SSE differences on the considered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

4.2. 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 Figure 1 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 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.

4.3. 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

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

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 (SSE) 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", defined by SSE of the 85th percentile. This was $\tau \approx 9$ for ISTA and $\tau \approx 20$ for VISTA. The middle panel shows the distribution of error for each as pair of box-plots 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 [19].

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.

Acknowledgements

The authors gratefully acknowledge support from the McCourt Institute and the Massive Data Institute. We thank Simon Segert for valuable conversations and input. Any errors remain our own.

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] Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Weighted nuclear norm minimization with application to image denoising. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2862–2869, 2014.

[12] 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.

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

[14] 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.

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

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

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

[18] 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.

[19] 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.

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

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

[22] 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.

[23] Nathan Wycoff, Ali Arab, Katharine M. Donato, and Lisa O. Singh. Sparse Bayesian lasso via a variable-coefficient ℓ_1 penalty, 2022. URL https://arxiv.org/abs/2211.05089.

[24] 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.

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