Precision Matrix Estimation with Noisy and Missing Data

Roger Fan[1] **Byoungwook Jang**[1] **Yuekai Sun**[1] **Shuheng Zhou**[2]

[1]University of Michigan
[2]University of California, Riverside

Abstract

Estimating conditional dependence graphs and precision matrices are some of the most common problems in modern statistics and machine learning. When data are fully observed, penalized maximum likelihood-type estimators have become standard tools for estimating graphical models under sparsity conditions. Extensions of these methods to more complex settings where data are contaminated with additive or multiplicative noise have been developed in recent years. In these settings, however, the relative performance of different methods is not well understood and algorithmic gaps still exist. In particular, in high-dimensional settings these methods require using non-positive semidefinite matrices as inputs, presenting novel optimization challenges. We develop an alternating direction method of multipliers (ADMM) algorithm for these problems, providing a feasible algorithm to estimate precision matrices with indefinite input and potentially nonconvex penalties. We compare this method with existing alternative solutions and empirically characterize the tradeoffs between them. Finally, we use this method to explore the networks among US senators estimated from voting records data.

1 Introduction

Undirected graphs are often used to describe high-dimensional distributions. Under sparsity conditions, these graphs can be estimated using penalized methods such as

$$\hat{\Theta} \in \underset{\Theta \succeq 0}{\arg\min} \left\{ \operatorname{tr}(\hat{\Gamma}_n \Theta) - \log\det(\Theta) + g_\lambda(\Theta) \right\}, \quad (1)$$

where $\hat{\Gamma}_n$ is the sample covariance or correlation matrix and g_λ is a separable (entry-wise) sparsity-inducing penalty function. Although this approach has proven successful in a variety of application areas such as neuroscience and genomics, its soundness hinges on the positive semidefiniteness (PSD) of $\hat{\Gamma}_n$. If $\hat{\Gamma}_n$ is indefinite, the objective may be unbounded from below.

In order to ensure this penalized M-estimator is well-behaving, Loh and Wainwright (2015) impose a side constraint of the form $\rho(\Theta) < R$, where ρ is a convex function. Here we focus on the estimator using the operator norm as a side constraint

$$\hat{\Theta} \in \underset{\Theta \succeq 0, \|\Theta\|_2 \leq R}{\arg\min} \left\{ \operatorname{tr}(\hat{\Gamma}_n \Theta) - \log\det(\Theta) + g_\lambda(\Theta) \right\}. \quad (2)$$

Loh and Wainwright (2017) adopt this method and show in theory the superior statistical properties of this constrained estimator. Their results suggest that the addition of a side constraint is not only sufficient but also almost necessary to effectively untangle the aforementioned complications.

Unfortunately, this additional constraint precludes using existing methods to solve the penalized objective with non-PSD input. To close this gap, we develop an alternating direction method of multipliers (ADMM) algorithm to implement (2) efficiently. We conduct empirical studies comparing this new method to several other precision matrix estimators. Our simulation study reveals several trends that are not present in the fully observed case. Finally, we illustrate the performance of our methods in analyzing the US senate voting data, uncovering both known and novel phenomena from the modern political landscape.

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

The remainder of this paper is organized as follows. In Section 2, we provide an overview of existing related work and describe in detail the optimization issues that arise from indefinite inputs and nonconvex penalties. In Section 3, we present the proposed ADMM algorithm and present some convergence results. Section 4 provides numerical examples and comparisons. Section 5 presents an exploratory analysis of US Senate voting records data using this method and details several interesting conclusions that can be drawn from the estimated graphs. Finally, we summarize the empirical results and their practical implications regarding choice of method in Section 6.

2 Problem formulation and existing work

There is a wide body of work proposing methods to perform precision matrix estimation in the fully observed case, including Meinshausen and Bühlmann (2006), Yuan and Lin (2007), Rothman et al. (2008), Friedman et al. (2008), Banerjee et al. (2008), and Zhou et al. (2010), most of which are essentially a ℓ_1-penalized likelihood approach (1) which we will refer to as the graphical Lasso.

Recent work has focused on using nonconvex regularizers such as SCAD and MCP for model selection in the regression setting (Fan and Li, 2001; Zhang, 2010; Breheny and Huang, 2011; Zhang and Zhang, 2012). Loh and Wainwright (2015, 2017) extend this analysis to general M-estimators, including variants of the graphical Lasso objective, and show their statistical convergence and support recovery properties. Estimators with these penalties have been shown to attain model selection under weaker theoretical conditions, but require more sophisticated optimization algorithms to solve, such as the local linear approximation (LLA) method of Fan et al. (2014).

In a fully observed and noiseless setting, $\hat{\Gamma}_n$ is the sample covariance and guaranteed to be at least positive semidefinite. Then, if g_λ is the ℓ_1-penalty, the objective of (1) is convex and bounded from below. In this setting, one can show that for $\lambda > 0$ a unique optimum $\hat{\Theta}$ exists with bounded eigenvalues and that the iterates for any descent algorithm will also have bounded eigenvalues (for example, see Lemma 2 in Hsieh et al., 2014).

When working with missing, corrupted, and dependent data, the likelihood is nonconvex, and the expectation-maximization (EM) algorithm has traditionally been used to perform statistical inference. However, in these noisy settings, the convergence of the EM algorithm is difficult to guarantee and is often slow in practice. For instance, Städler and Bühlmann (2012) implement a likelihood-based method for inverse covariance estimation with missing values, but their EM algorithm requires solving a full graphical Lasso optimization problem in each M-step.

An alternative approach is to develop M-estimators that account for missing and corrupted data. For graphical models, Loh and Wainwright (2015) establish that the graphical Lasso, including a version using nonconvex penalties, can be modified to accommodate noisy or missing data by adjusting the sample covariance estimate.

These modified estimators depend on the observation that statistical theory for the graphical Lasso generally requires that $\|\hat{\Gamma}_n - \Sigma\|_\infty$ converges to zero at a sufficiently fast rate (e.g. Rothman et al., 2008; Zhou et al., 2010; Loh and Wainwright, 2017). When considering missing or corrupted data, it is often possible to construct covariance estimates $\hat{\Gamma}_n$ that satisfy this convergence criteria but are not necessarily positive semidefinite. In fact, in high-dimensional settings $\hat{\Gamma}_n$ may even be guaranteed to be indefinite. Attempting to input these indefinite covariance estimates into the graphical Lasso, however, presents novel optimization issues.

Unbounded objective. When attempting to move beyond the ℓ_1 penalized case with positive semidefinite input, the problem in (1) becomes unbounded from below, so an optimum may not necessarily exist. This issue comes from two potential sources: 1) negative eigenvalues in $\hat{\Gamma}_n$, or 2) zero eigenvalues combined with the boundedness of the nonconvex penalty g_λ. For example, consider the restriction of the objective in (1) to a ray defined by an eigenvalue-vector pair σ_1, v_1 of $\hat{\Gamma}_n$:

$$\begin{aligned} &f(I + tv_1v_1^T) \\ &\quad = \operatorname{tr}(\hat{\Gamma}_n) + t\operatorname{tr}(\hat{\Gamma}_n v_1 v_1^T) - \log(1+t) + g_\lambda(tv_1v_1^T) \\ &\quad = \operatorname{tr}(\hat{\Gamma}_n) + t\sigma_1 - \log(1+t) + g_\lambda(tv_1v_1^T). \end{aligned} \tag{3}$$

If $\sigma_1 < 0$, we see that f is unbounded from below due to the $t\sigma_1$ and $-\log(1+t)$ terms. In fact, if $\sigma_1 = 0$ and g_λ is bounded from above, as is the case when using standard nonconvex penalties, the objective is also unbounded from below.

So unboundedness can occur anytime there is a negative eigenvalue in the input matrix, or whenever there are zero eigenvalues combined with a nonconvex penalty function g_λ. Unboundedness creates optimization issues, as an optimum no longer necessarily

exists.

Handling unboundedness. In order to guarantee that an optimum exists for (1), an additional constraint of the form $\rho(\Theta) \leq R$ can be imposed, where ρ is some convex function. In this paper, we consider the estimator (2), which uses a side constraint of the form $\|\Theta\|_2 \leq R$. Loh and Wainwright (2017) show the rates of convergence of this estimator (2) and show that it can attain model selection consistency and spectral norm convergence without the incoherence assumption when used with a nonconvex penalty (see Appendix E therein), but do not discuss implementation or optimization aspects of the problem.

To our knowledge, there is currently no feasible optimization algorithm for the estimator defined in (2), particularly when the input is indefinite. Loh and Wainwright (2015) present a composite gradient descent method for optimizing a subset of side-constrained versions of (1). However, their algorithm requires a side constraint of the form $\rho(\Theta) = \frac{1}{\lambda}(g_\lambda(\Theta) + \frac{\mu}{2}\|\Theta\|_F^2)$, which does not include the spectral norm constraint and therefore cannot attain the better theoretical results it achieves (Section C.7 compares the performance of different side constraints). It may be possible to develop heuristic algorithms that alternate performing a proximal gradient update ignoring the side constraint and projecting to the constraint set, but as far as we know there has not been any analysis of algorithms of this type (we discuss this in more detail in Section C.4).

An alternative approach to solving this unbounded issue is to project the input matrix $\hat{\Gamma}_n$ to the positive semidefinite cone before inputting into (1). We discuss this further in Section 4.1, but this only solves the unbounded issue when using the ℓ_1 penalty; nonconvex penalties still require a side constraint to have a bounded objective and therefore our algorithm is still useful even for the projected methods.

3 ADMM Algorithm

Our algorithm is similar to the algorithm in Guo and Zhang (2017), which applies ADMM to the closely related problem of condition number-constrained sparse precision matrix estimation using the same splitting scheme as below. We discuss their method in more detail in Section A.6. The following algorithm is specialized to the case where the spectral norm is used as the side constraint. In Section B we derive a similar ADMM algorithm that can be used for any side constraint with a computable projection operator.

Algorithm 1: ADMM for graphical Lasso with a side constraint

Input: $\hat{\Gamma}_n$, ρ, g_λ, R
Output: $\hat{\Theta}$
Initialize $V^0 = \Theta^0 \succ 0$, $\Lambda^0 = \mathbf{0}$;
while *not converged* **do**
$V^{k+1} = \text{Prox}_{g_\lambda/\rho}\left(\frac{\rho\Theta^k + \Lambda^k}{\rho}\right)$
$\Theta^{k+1} = T_\rho\left(\frac{\rho V^{k+1} - \hat{\Gamma}_n - \Lambda^k}{\rho}\right)$
$\Lambda^{k+1} = \Lambda^k + \rho(\Theta^{k+1} - V^{k+1})$
end

Rewrite the objective from (2) as

$$f(\Theta) = \text{tr}(\hat{\Gamma}_n\Theta) - \log\det(\Theta) + g_\lambda(\Theta) + \mathbb{1}_{\mathcal{X}_R}(\Theta) \quad (4)$$

where $\mathcal{X}_R = \{\Theta : \Theta \succeq 0, \|\Theta\|_2 \leq R\}$ and $\mathbb{1}_{\mathcal{X}}(\Theta) = 0$ if $\Theta \in \mathcal{X}$ and ∞ otherwise.

Let $\rho > 0$ be a penalty parameter and let $\text{Prox}_{g_\lambda/\rho}$ be the prox operator of g_λ/ρ. We derive these updates for SCAD and MCP in Section A.2. Let $T_\rho(A)$ be the following prox operator for $-\log\det\Theta + \mathbb{1}_{\mathcal{X}_R}(\Theta)$, which we derive in Section A.3,

$$T_\rho(A) = T_\rho(UMU^T) = U\tilde{D}U^T$$
$$\text{where } \tilde{D}_{ii} = \min\left\{\frac{M_{ii} + (M_{ii}^2 + \frac{4}{\rho})^{1/2}}{2}, R\right\},$$

where UMU^T is the eigendecomposition of A. Then the ADMM algorithm for solving (4), which we derive in Section A.2, is described in Algorithm 1. Computationally this algorithm is dominated by the eigendecomposition used to evaluate T_ρ, and therefore has a complexity of $O(m^3)$, which matches the scaling of other graphical Lasso solvers (e.g. Meinshausen and Bühlmann, 2006; Friedman et al., 2008; Hsieh et al., 2014).

3.1 Convergence

The following proposition applies standard results on the convergence of ADMM for convex problems to show convergence when the ℓ_1 penalty is used. Details are in Section A.4.

Proposition 1. *If the penalty is convex and satisfies the conditions in Section A.1, Algorithm 1 converges to a global minimum of* (4).

Remark. Regarding the nonconvex penalty, recent work has established ADMM convergence results in some nonconvex settings (see Hong et al., 2016; Wang et al., 2015), but to our knowledge there is no convergence result that encompasses this nonsmooth and

nonconvex application. We can show convergence if a fairly strong assumption is made on the iterates, but we are currently working on extending existing results to this case.

Proposition 2 shows that any limiting point of Algorithm 1 is a stationary point of the original objective (4). This is proved in Section A.5. When using the ℓ_1 penalty or a nonconvex penalty with $R \leq \sqrt{2/\mu}$, where μ is the weak convexity constant of g_λ, the objective f is convex and therefore any stationary point is unique and also the global optimum. See Section C.5 for a more detailed discussion.

Proposition 2. *Assume that the penalty g_λ satisfies the conditions in Section A.1. Then for any limit point $(\Theta^*, V^*, \Lambda^*)$ of the ADMM algorithm defined in Algorithm 1, Θ^* is also a stationary point of the objective f as defined in* (4).

The assumptions on g_λ in Section A.1 are the same as those assumed in Loh and Wainwright (2015, 2017), and are satisfied by the Lasso, SCAD, and MCP functions.

Note that if a limiting point is found to exist when using a nonconvex penalty the result in Proposition 2 will still hold. Empirically we find that the algorithm performs well and converges consistently when used with nonconvex penalties, but there is no existing theoretical guarantee that a limiting point of ADMM will exist in that setting.

4 Simulations

We evaluate the proposed estimators using the relative Frobenius norm and the sum of the false positive rate and false negative rate (FPR+FNR). We present results over a range of λ values, noting that all the compared methods would use similar techniques to perform model tuning. Section C.1 presents an example of how to use BIC or cross-validation to tune these methods. We present results using covariance matrices from auto-regressive and Erdős-Rényi random graph models. See Section C for descriptions of these models as well as additional simulation results.

4.1 Alternative methods

When faced with indefinite input, there are two alternative graphical Lasso-style estimators that can be used besides (2), which involve either ℓ_∞ projection to the positive semidefinite cone or nodewise regression in the style of Meinshausen and Bühlmann (2006).

Projection. Given an indefinite input matrix $\hat{\Gamma}_n$, Park (2016) and Greenewald et al. (2017) propose performing the projection $\hat{\Gamma}_n^+ = \arg\min_{\Gamma \succeq 0} \|\Gamma - \hat{\Gamma}_n\|_\infty$. They then input $\hat{\Gamma}_n^+$ into the optimization problem (1). This is similar to the projection done in Datta and Zou (2017). In terms of the upper bound on statistical convergence rates, this method pays a constant factor cost, though in practice projection may result in a loss of information and therefore a decrease in efficiency.

After projecting the input, existing algorithms can be used to optimize (1) with the ℓ_1 penalty. However, as mentioned in Section 2, using a nonconvex penalty still leads to an unbounded objective and therefore still requires using our ADMM algorithm to solve (2).

Nodewise regression. Loh and Wainwright (2012) and Rudelson and Zhou (2017) both study the statistical and computational convergence properties of using errors-in-variables regression to handle indefinite input matrices in high-dimensional settings. Following the nodewise regression ideas of Meinshausen and Bühlmann (2006) and Yuan (2010), we can perform m Lasso-type regressions to obtain estimates $\hat{\beta}_j$ and form estimates $\hat{a}_j$, where

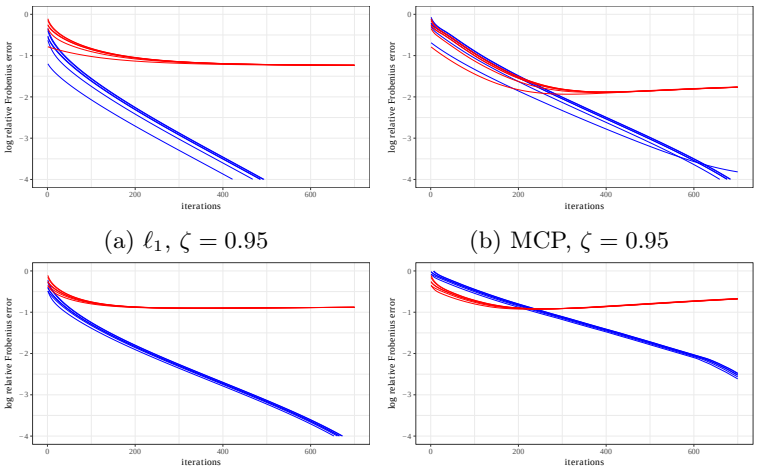
$$\begin{aligned} \hat{\beta}_j &\in \underset{\|\beta\|_1 \leq R}{\arg\min} \left\{ \frac{1}{2} \beta^T \hat{\Gamma}_{n,-j,-j} \beta - \langle \hat{\Gamma}_{n,-j,j}, \beta \rangle + \|\beta\|_1 \right\} \\ \hat{a}_j &= -(\hat{\Gamma}_{n,j,j} - \langle \hat{\Gamma}_{n,-j,j}, \hat{\beta}_j \rangle)^{-1} \end{aligned} \tag{5}$$

and combine to get $\tilde{\Theta}$ with $\tilde{\Theta}_{-j,j} = \hat{a}_j \hat{\beta}_j$ and $\tilde{\Theta}_{j,j} = -\hat{a_j}$. Finally, we symmetrize the result to obtain $\hat{\Theta} = \arg\min_{\Theta \in S^m} \|\Theta - \tilde{\Theta}\|_1$, where S^m is the set of symmetric matrices.

These types of nodewise estimators have gained popularity as they require less restrictive incoherence conditions to attain model selection consistency and often perform better in practice in the fully observed case. They have not, however, been as well studied when used with indefinite input.

4.2 Data models

We test these methods on two models that result in indefinite covariance estimators, the non-separable Kronecker sum model from Rudelson and Zhou (2017) and the missing data graphical model described in Loh and Wainwright (2015). In the main paper we focus on the missing data model, but Section C contains a detailed description of the Kronecker sum model as well as simulation results using it.



(a) ℓ_1, $\zeta = 0.95$ (b) MCP, $\zeta = 0.95$

(c) ℓ_1, $\zeta = 0.7$ (d) MCP, $\zeta = 0.7$

Figure 1: Convergence of the ADMM algorithm for several initializations. Blue lines show the relative optimization error ($\|\Theta^k - \hat{\Theta}\|_F / \|\Theta^*\|_F$, where $\hat{\Theta}$ is the result of running our algorithm to convergence) while red lines show the statistical error ($\|\Theta^k - \Theta^*\|_F / \|\Theta^*\|_F$). All panels use an AR1(0.7) covariance with $m = 300$ and $n = 125$ and set $\rho = 12$. The left panels use an ℓ_1 penalty, while the right panels use MCP with $a = 2.5$. R is set to be three times the oracle spectral norm.

Missing data (MD). As discussed above, Loh and Wainwright (2013, 2015) propose an estimator for a graphical model with missing-completely-at-random observations.

Let $W \in \mathbb{R}^{n \times m}$ be a mean-zero subgaussian random matrix. Let $U \in \{0,1\}^{n \times m}$ where $U_{ij} \sim$ Bernoulli(ζ_j) are independent of W. This corresponds to entries of the jth column of the data matrix being observed with probability ζ_j. Then we have an unobserved matrix Z and observed matrix X generated by $Z = WA^{1/2}$ and $X = U \circ X$, where $\circ$ denotes the Hadamard, or element-wise, product. Here the covariance estimate for A is

$$\hat{\Gamma}_n = \frac{1}{n} X^T X \oslash M \text{ where } M_{k\ell} = \begin{cases} \zeta_k & \text{if } k = \ell \\ \zeta_k \zeta_\ell & \text{if } k \neq \ell \end{cases} \tag{6}$$

where $\oslash$ denotes element-wise division. As we divide off-diagonal entries by smaller values, $\hat{\Gamma}_n$ will not necessarily be positive semidefinite.

4.3 Simulation results

Optimization performance. Figure 1 shows the optimization performance of Algorithm 1 using nonprojected input matrices from the missing data model with both ℓ_1 and nonconvex penalties (MCP). The top two panels present an "easy" scenario with a higher sampling rate, while the bottom two have a more challenging scenario with significant missing data. Blue lines report the optimization error while red lines are the statistical error.

All the plots in Figure 1 have their optimization error quickly converge to below the statistical error. These plots also suggest that our algorithm can attain linear convergence rates. We find that the algorithm consistently converges well over a range of tested scenarios.

Comparing the statistical error of the top two plots, we see that MCP achieves significantly lower error for the easier scenario. But in the bottom two plots, where there is more missing data, it struggles relative to the ℓ_1 penalty. This is a common trend through our simulations, as the performance of estimators using MCP degrades as missingness increases while the ℓ_1-penalized versions are more robust.

Method comparisons. Figure 2 demonstrates the statistical performance along the full regularization path. Across the panels from left to right, the sampling rate decreases and therefore the magnitude of the most negative eigenvalue increases (see Table 4).

In terms of Frobenius error, both projected methods and the nonprojected estimator with the ℓ_1 penalty get slightly worse across panels, but the nodewise regression and the nonprojected MCP estimator react much more negatively to more indefinite input. The nodewise regression in particular goes from being among the best to among the worst estimators as the sampling rate decreases.

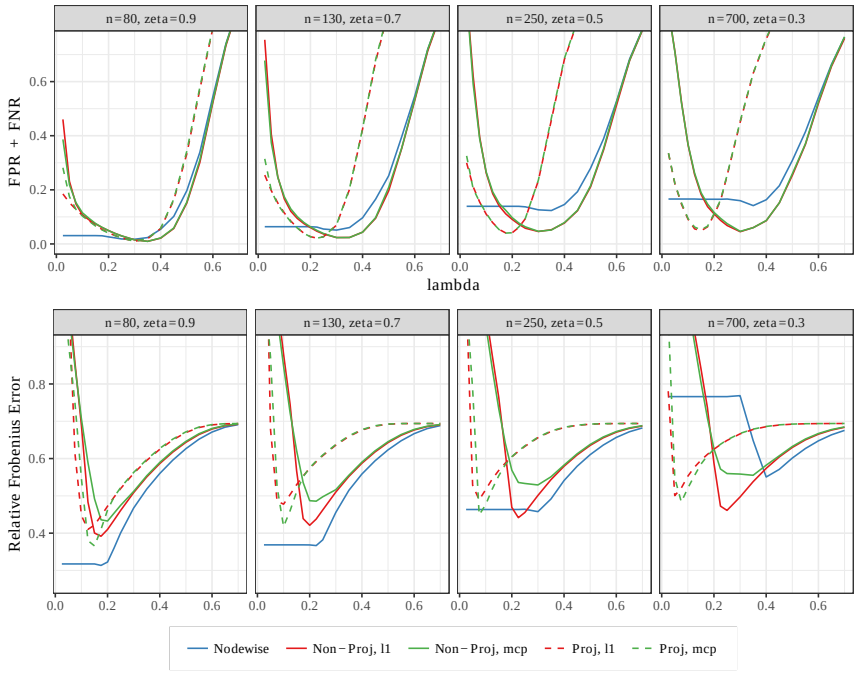


Figure 2: The performance of the various estimators for the missing data model in terms of relative Frobenius error ($\|\hat{\Theta} - \Theta^*\|_F / \|\Theta^*\|_F$) and model selection as measured by FPR + FNR. We use an AR(0.6) covariance and set $m = 1200$. Settings are chosen so that the effective sample size ($n\zeta^2$) is roughly equivalent. The MCP penalty uses $a = 2.5$. We set R to be 1.5 times the oracle value for each method and set $\rho = 24$. Our convergence criteria is $\|\Theta^{k+1} - \Theta^k\|_F / \|\Theta^k\|_F < 5\text{e}{-5}$.

Comparing the projected and nonprojected curves in Figure 2, we see that the optimal value of λ, as well as the range of optimal values, shrinks for the projected method as the sampling rate decreases. This pattern is consistently repeated across models and scenarios, likely because the ℓ_∞ projection is shrinking the off-diagonal entries of the input matrix. We find that the nonprojected graphical Lasso performs slightly better than the projected version when used with the ℓ_1 penalty, likely due to the information lost in this shrinkage.

Figure 2 also shows how these methods perform in terms of model selection. We can see that the nonconvex penalties perform essentially identically to their ℓ_1 penalized counterparts. In particular, the degradation of the nonprojected MCP estimator in terms of norm error does not seem to affect its model selection performance. The nodewise regression, however, still demonstrates this pattern, as its model selection performance degrades across the panels. For scenarios with more missing data, the nonprojected estimators seem to be easier to tune, maintaining a wider range of λ values where they perform near-optimally. In Section C of the supplement we perform similar experiments in a variety of different noise and model settings.

Sensitivity to R. Figure 3 demonstrates the sensitivity of the nonprojected estimators to the choice of R, the size of the side constraint. We can see that all these methods are sensitive to the choice of R for small values of λ in terms of norm error. None of the methods are sensitive in terms of model selection.

The nonprojected graphical Lasso with MCP is the most sensitive to R and is also sensitive for larger choices of λ, which is important since it never reaches its oracle minimum norm errors when R is chosen to be larger than the oracle. The nonprojected graphical Lasso with ℓ_1 and the projected graphical Lasso with MCP both still achieve the same best-case performance when R is misspecified, though tuning λ becomes more difficult.

The nodewise regression results are also plotted here. Here R is the ℓ_1 side constraint level in (5). For smaller values of λ the nodewise estimator levels off, corresponding to when the side constraint becomes active over the penalty. Different values of R change when this occurs and, if R is chosen large enough, do not significantly affect ideal performance. Note that these use a stronger oracle that knows each column-

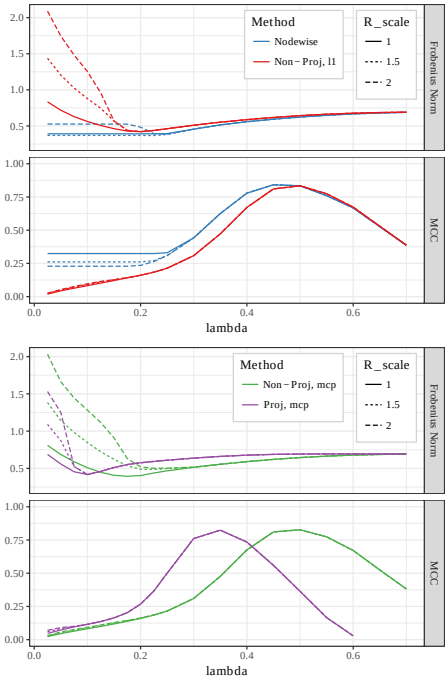


Figure 3: The performance of missing data estimators over different choices of R. The non-nodewise estimators set $R = \text{R\_scale} \times \|A\|_2$, while each node's regression in the nodewise estimator sets R to be R_scale times that node's oracle ℓ_1 value. We use an $\text{AR}(0.6)$ covariance, set $m = 1200$, $n = 130$, and choose a sampling rate of $\zeta = 0.7$. The MCP penalty is chosen with $a = 2.5$.

wise ℓ_1 norm, but do show that this method can be improved with careful tuning.

5 Senate voting analysis

Based on the missing data model from Section 4.2, we estimate the conditional dependence graph among senators using the ADMM algorithm from Section 3. The dataset includes voting records from the United States Senate during the 112th Congress (2011-2013). We drop senators who serve partial terms and unanimous votes, resulting in a dataset of voting records for 99 senators over 426 votes. Appendix D contains further details regarding data processing and the methods used as well as additional analysis.

Missing values in this data correspond to votes that are missed by senators and consist of roughly 2.6% of total votes. Note that only 109 of the votes are fully observed, so some type of correction or imputation should be used instead of omitting rows.

A major story at this time was the rise of the tea party movement in the Republican party. Across the US government tea party challengers rose to prominence. Though it was not an official party, politicians associated with the tea party movement tended to be more conservative and less likely to compromise than establishment Republicans, leading to a particularly politically polarized period of government.

Figure 4 plots the estimated graph among senators. As expected the distinction between Republicans and Democrats is stark. Both independent senators caucus with the Democrats, so as expected they are part of the Democratic component of the graph.

We identify senators who were present at the inaugural meeting of the unofficial Senate Tea Party Caucus as well as those elected in 2010 with significant tea party support.[1] These senators are colored in black, and we can see that within the Republican party they are clustered together.

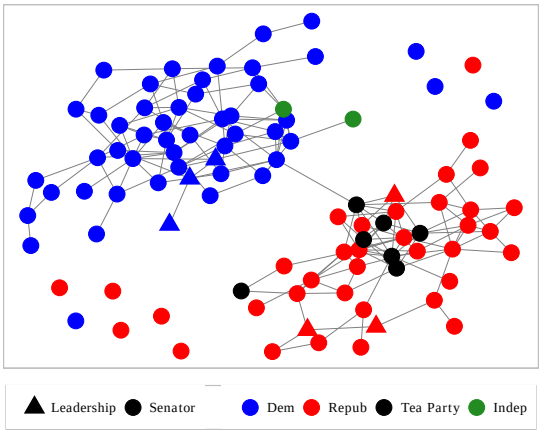
In Figure 4a we can see that the sole connection between parties runs through the tea party (Rand Paul) and Jeff Merkley, a Democratic senator. This may be surprising, as Rand Paul is one of the most conservative senators and Merkley one of the most liberal. Paul is, however, regarded as a relatively libertarian conservative. So though he is extremely conservative in some dimensions, he may share liberal views with Merkeley on others.

Figure 4b plots the same graph estimated at a lower penalization level. The Republicans who have cross-party connections include some of both the most conservative (Paul) and the most moderate (Thad Cochran, Lisa Murkowski).[2] On the Democratic side the cross-connected senators also include both the most liberal (Sanders, Merkley, Tom Udall) and relatively moderate (Claire McCaskill). As expected, moderates are among those most connected opposing party, but this shows that the most extreme members of a party can also be linked to the opposing party. Appendix D discusses these cross-party links in more detail.

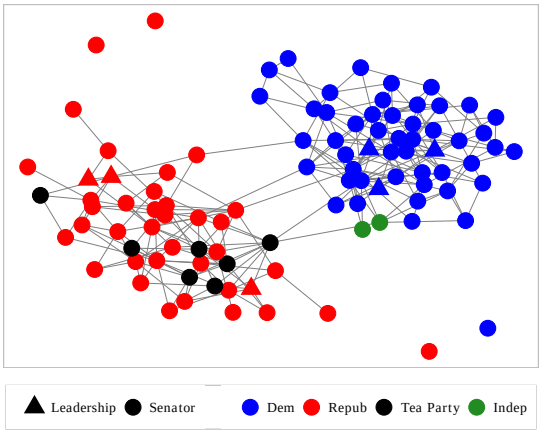
Figure 4c shows the Republican subgraph from Figure 4a. Here we can identify other senators who are closely associated with the tea party. In particular, two nodes near the tea party cluster are marked 'H' and 'C,' corresponding to Senators Orrin Hatch and Tom Coburn. Both have been linked to the tea party in the media, either as candidates supported by it or

[1] The marked tea party senators are Marco Rubio, Mike Lee, Jerry Moran, Jim DeMint, Rand Paul, Ron Johnson, and Pat Toomey.

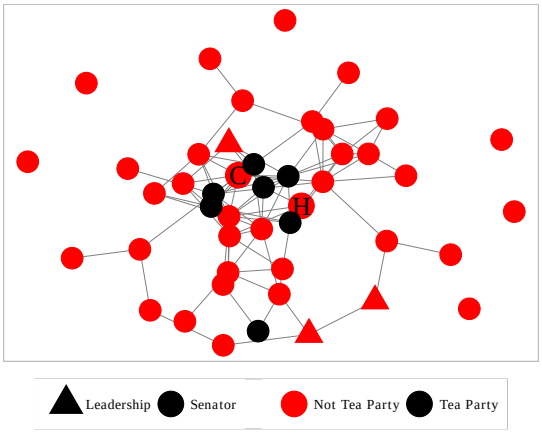
[2] Here we are measuring ideology by NOMINATE, a standard method in political science for assessing a representative's position on the political spectrum (Poole, 2005). See Appendix D for more details.



(a) $\lambda = 0.21$



(b) $\lambda = 0.15$



(c) $\lambda = 0.15$, Republican subgraph

Figure 4: Graphs among senators estimated on Senate voting records from the 112th US Congress using an ℓ_1 penalty with penalty λ as indicated. We set $R = 10$ and the ADMM algorithm was run with $\rho = 10$. After estimation, the precision matrix is thresholded at 0.04 for the top panel and 0.055 for the bottom two.

as being supportive of the movement.

It is also of interest that one marked senator is not clustered with the others, Jerry Moran. This suggests that he is not as closely connected to the tea party movement as the others we have identified.

6 Summary and discussion

In this paper, we study the estimation of sparse precision matrices from noisy and missing data. To close an existing algorithmic gap, we propose an ADMM algorithm that allows for fast optimization of the side-constrained graphical Lasso, which is needed to implement the graphical Lasso with either indefinite input and/or nonconvex penalties. We investigate its convergence properties and compare its performance with other methods that handle the indefinite sample covariance matrices that arise with dirty data.

We find that methods with nonconvex penalties are quite sensitive to the indefiniteness of the input covariance estimate, and are particularly sensitive to the magnitude of its negative eigenvalues. They may have better existing theoretical guarantees, but in practice we find that with nontrivial missingness or noise they perform worst than or, at best, recover the performance of their ℓ_1-normalized counterparts. The nonconvex methods can outperform the ℓ_1-penalized ones when there is a small amount of missingness or noise, but in these cases we often find the nodewise estimator to perform best.

In difficult settings with significant noise or missingness, the most robust and efficient method seems to be using the graphical Lasso with nonprojected input and an ℓ_1 penalty. As the application becomes easier – with more observations or less missing data – the nodewise estimator becomes more competitive, just as it is understood to be with fully observed data.

The projected graphical Lasso estimator with an ℓ_1 penalty seems to be slightly worse than its nonprojected counterpart. Projection does, however, allow for the use of nonconvex penalties in more difficult settings without the large degradation in performance we have observed. This may be desired in some scenarios when the nonzero off-diagonal precision matrix entries are expected to be large.

Finally, we also use this new algorithm to estimate conditional dependence graphs among US senators using voting records data. We identify several interesting patterns in these graphs, especially regarding the rise of the tea party movement and cross-party connections between senators.

Acknowledgements

The research is supported in part by the NSF under grants DMS-1316731 and NSF-1830247.

References

Agarwal, A., Negahban, S., and Wainwright, M. J. (2012). Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, 40(5):2452–2482.

Banerjee, O., Ghaoui, L. E., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9(Mar):485–516.

Belloni, A., Rosenbaum, M., and Tsybakov, A. B. (2017). Linear and conic programming estimators in high dimensional errors-in-variables models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):939–956.

Bertsekas, D. P. and Tsitsiklis, J. N. (1989). *Parallel and distributed computation: Numerical methods*. Prentice Hall Englewood Cliffs, NJ. Republished by Athena Scientific in 1997.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122.

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, 5(1):232.

Datta, A. and Zou, H. (2017). Cocolasso for high-dimensional error-in-variables regression. *The Annals of Statistics*, 45(6):2400–2426.

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

Fan, J., Xue, L., and Zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Annals of Statistics*, 42(3):819.

Farrell, R. H. (1985). Multivariate calculation: Use of the continuous groups.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Greenewald, K., Park, S., Zhou, S., and Giessing, A. (2017). Time-dependent spatially varying graphical models, with application to brain fMRI data analysis. In *Advances in Neural Information Processing Systems*, pages 5834–5842.

Guo, X. and Zhang, C. (2017). The effect of L_1 penalization on condition number constrained estimation of precision matrix. *Statistica Sinica*, 27:1299–1317.

Hong, M., Luo, Z.-Q., and Razaviyayn, M. (2016). Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. *SIAM Journal on Optimization*, 26(1):337–364.

Hornstein, M., Fan, R., Shedden, K., and Zhou, S. (2018). Joint mean and covariance estimation with unreplicated matrix-variate data. *Journal of the American Statistical Association*.

Hsieh, C.-J., Sustik, M. A., Dhillon, I. S., and Ravikumar, P. (2014). QUIC: Quadratic approximation for sparse inverse covariance estimation. *Journal of Machine Learning Research*, 15(1):2911–2947.

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664.

Loh, P.-L. and Wainwright, M. J. (2013). Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *The Annals of Statistics*, 41(6):3022.

Loh, P.-L. and Wainwright, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16:559–616.

Loh, P.-L. and Wainwright, M. J. (2017). Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics*, 45(6):2455–2482.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462.

Mota, J. F., Xavier, J. M., Aguiar, P. M., and Püschel, M. (2011). A proof of convergence for the alternating direction method of multipliers applied to polyhedral-constrained functions. *arXiv preprint arXiv:1112.2295*.

Park, S. (2016). *Selected Problems for High-Dimensional Data–Quantile and Errors-in-Variables Regressions*. PhD thesis, University of Michigan.

Park, S., Shedden, K., and Zhou, S. (2017). Nonseparable covariance models for spatio-temporal

data, with applications to neural encoding analysis. *arXiv preprint arXiv:1705.05265*.

Poole, K. T. (2005). *Spatial models of parliamentary voting*. Cambridge University Press.

Rosenbaum, M. and Tsybakov, A. B. (2010). Sparse recovery under matrix uncertainty. *The Annals of Statistics*, pages 2620–2651.

Rosenbaum, M. and Tsybakov, A. B. (2013). Improved matrix uncertainty selector. In *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, pages 276–290. Institute of Mathematical Statistics.

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515.

Rudelson, M. and Zhou, S. (2017). Errors-in-variables models with dependent measurements. *Electronic Journal of Statistics*, 11(1):1699–1797.

Städler, N. and Bühlmann, P. (2012). Missing values: sparse inverse covariance estimation and an extension to sparse regression. *Statistics and Computing*, 22(1):219–235.

Wang, Y., Yin, W., and Zeng, J. (2015). Global convergence of ADMM in nonconvex nonsmooth optimization. *arXiv preprint arXiv:1511.06324*.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11(Aug):2261–2286.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35.

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

Zhang, C.-H. and Zhang, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, pages 576–593.

Zhou, S. (2014). GEMINI: Graph estimation with matrix variate normal instances. *The Annals of Statistics*, 42(2):532–562.

Zhou, S. (Forthcoming, 2019). Sparse Hanson-Wright inequalities for subgaussian quadratic forms. *Bernoulli*. Available at https://arxiv.org/abs/1510.05517.

Zhou, S., Lafferty, J., and Wasserman, L. (2010). Time varying undirected graphs. *Machine Learning*, 80:295–319.