Robust Estimation of Laplacian Constrained Gaussian Graphical Models with Trimmed Non-convex Regularization

Mariana Vargas Vieyra

Robust Estimation of Laplacian Constrained Gaussian Graphical Models with Trimmed Non-convex Regularization

Mariana Vargas Vieyra [1]

Abstract

The problem of discovering a structure that fits a collection of vector data is of crucial importance for a variety of applications. Such problems can be framed as Laplacian constrained Gaussian Graphical Model inference. Existing algorithms rely on the assumption that all the available observations are drawn from the same Multivariate Gaussian distribution. However, in practice it is common to find scenarios where the datasets are contaminated with a certain number of outliers. The purpose of this work is to address that problem. We propose a robust method based on Trimmed Least Squares that copes with the presence of corrupted samples. We provide statistical guarantees on the estimation error and present results on simulated data.

1. Introduction

Numerous modern applications rely on relational information to study a particular phenomenon or address a specific task. For example, in social network analysis, predictions about the behavior of a particular user may also depend on the behavior of its mutual friends. The need to incorporate relational information into Machine Learning algorithms calls for data processing methods that can handle highly correlated data. In this context, graphs have been widely used because of their ability to compactly encode pairwise relationships between points. While in many domains the graph structure arises naturally, in many scenarios it is often the case that no *a-priori* graph is readily available. The problem of estimating the graph structure that fits a collection of data can be framed as Laplacian Constrained Gaussian Graphical Model (LCGGM) inference.

In its general form, a Gaussian Graphical Model (GGM) is parameterized by a matrix called the *precision matrix* that encodes the conditional dependence relationships between variables. Most methods for estimating GGMs impose an L^1 penalization on the parameters to encourage sparisity. This is known as Graphical Lasso (Friedman et al., 2008; Banerjee et al., 2005; Cai et al., 2016). When constrained to graph Laplacian matrices, the precision matrix of the GGM can be deemed an operator with specific properties that permits to reason about the observed data. In particular, the eigenvectors of a Laplacian matrix provide a Fourier basis of the data, and the eigenvalues can be interpreted as spectral frequencies (von Luxburg, 2007). The study of graph Laplacians have led to the field of Graph Signal Processing (Ortega et al., 2018; Shuman et al., 2013; Dong et al., 2020).

In general, methods for LCGGM estimation encourage sparse estimators (Ying et al., 2020; 2021; Lake & Tenenbaum, 2010; Koyakumaru et al., 2021), although guarantees for non-regularized Maximum Likelihood Estimators have been explored (Egilmez et al., 2017; Ying et al., 2021; Pavez, 2022). Other algorithms take a different route and propose estimators based on global smoothness (Dong et al., 2016; Kalofolias). Finally, algorithms that allow more restrictive constraints have been studied (Kumar et al., 2019). These approaches rely on the hypothesis that all observed elements come from the same distribution. However, in practice this is often a strong assumption, as available samples might be contaminated with outliers. While robust methods have been proposed in the context of vanilla GGMs (Yang & Lozano, 2015), the problem of estimating LCGGM with corrupted samples remains under explored.

In this work we propose a robust method based on Trimmed Least Squares to estimate sparse LCGGMs in scenarios where the available dataset has a relatively small proportion of outliers. We provide statistical guarantees on the estimation error and empirically validate our method on simulated data.

2. Background and Setting

We start by introducing the necessary notation and background on Laplacian matrices and Laplacian-constrained Graphical Models. We then describe the problem setup.

[1]Inria Grenoble, France. Correspondence to: Mariana Vargas Vieyra <mariana.vargas-vieyra@inria.fr>.

Presented at ICML Workshop on Principles of Distribution Shift 2022,

2.1. Graph Laplacian

Let us consider a graph $\mathcal{G} = (V, E, W)$ where V is a set of p nodes, $E \subseteq V \times V$ is a set of edges, and W is a matrix with non-negative entries such that $W_{ij} > 0$ if and only if $(v_i, v_j) \in E$. The graph Laplacian is defined as $L = D - W$, where D is a diagonal matrix with elements $D_{ii} = \sum_{j=1}^{p} W_{ij}$. If we consider graph Laplacians associated to graphs with one connected component, then the set of valid graph Laplacians can be defined as:

$$\mathcal{S}_L = \{\Theta : \Theta_{ij} \leq 0 \text{ if } i \neq j, \Theta \mathbf{1} = \mathbf{0}\} \tag{1}$$

The graph Laplacian is a crucial tool to reason about the structure of the graph and the data associated to it as they allow to generalize notions of Signal Processing to non-Euclidean, irregular domains (Bronstein et al., 2017). In particular, graph Laplacians provide a Fourier basis of the associated data through their spectral decomposition. They also provide a quadratic form that quantifies the smoothness of the data. Global smoothness has been widely used as a criterion to assess the quality of an estimated graph (Kalofolias; Dong et al., 2016; Tenenbaum et al., 2000).

2.2. Problem setup

Let $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_p]$ be a random vector with a Gaussian distribution, $\mathbf{X} \sim \mathcal{N}(0, \Sigma)$, and let $X = [x_1^\intercal, \ldots, x_n^\intercal]$ be a design matrix consisting of n p-dimensional observations of $\mathbf{X}$. The precision matrix $\Theta = \Sigma^{-1}$ characterizes the conditional independence structure of the variables. That is, $\mathbf{X}$ is associated with a graph $\mathcal{G}$ where the nodes correspond to the elements of $\mathbf{X}$ and $\Theta_{ij} > 0$ if and only if $\mathbf{X}_i$ and $\mathbf{X}_j$ are connected in the graph. Let $S = \frac{1}{n} X^\intercal X$ be the sample covariance matrix obtained from the data, and let J be a constant matrix defined such that $J_{ij} = 1/p$ for all i, j. We define the linear operator $\mathcal{L} : \mathbb{R}^{p(p-1)/2} \to \mathbb{R}^{p \times p}$ that maps vectors to Laplacian matrices introduced by Kumar et al. (2019) as

$$(\mathcal{L}w)_{ij} = \begin{cases} -w_k & i > j \\ (\mathcal{L}w)_{ji} & i < j \\ -\sum_{j \neq i} (\mathcal{L}w)_{ij} & i = j \end{cases}$$

where $k = i - j + \frac{j-1}{2}(2p - j)$. $\mathcal{L}$ has an adjoint operator $\mathcal{L}^* : \mathbb{R}^{p \times p} \to \mathbb{R}^{p(p-1)/2}$ defined as

$$(\mathcal{L}^* W)_i = W_{ii} - W_{ij} - W_{ji} + W_{jj} \,.$$

In this work we adopt the setting for learning Gaussian Graphical Models under Laplacian constraints proposed by Ying et al. (2020). That is,

$$\min_{w \geq 0} -\log\det(\mathcal{L}w + J) + \operatorname{Tr}(\mathcal{L}wS) + \sum_i r_\lambda(w_i) \tag{2}$$

where $\sum_i r_\lambda(w_i)$ is a nonconvex regularization function that encourages sparsity, and where we write $w \geq 0$ for element-wise comparison. Note that by adding J in the first term the argument of the log determinant becomes full rank (Egilmez et al., 2017).

We now assume that our data set is contaminated with samples drawn from a different distribution. This way, "good" samples come from the distribution with true parameter $\Theta^\star$ and "bad" samples come from a different distribution. Let $[n] = \{1, \ldots, n\}$. Let $G = \{i : X^{(i)} \sim \mathcal{N}(0, (\Theta^\star)^{-1})\}$ be the set of indexes of "good" samples and $B = [n] \setminus G$ the set of indexes of "bad" samples. We further denote $g = |G|$ and $b = |B|$ the number of "good" and "bad" samples respectively.

In the following we will introduce a method for robust estimation of Laplacian constrained graphical models and derive statistical guarantees on the estimation error.

3. Robust Estimation of Laplacian Constrained Gaussian Graphical Models

In this Section we first introduce a method inspired on Trimmed Least Squares for robust maximum likelihood estimation of Laplacian constrained graphical models with a nonconvex penalization. We then derive statistical guarantees for the estimation error.

3.1. Proposed Method

Following the same route as Yang & Lozano (2015), we propose to solve the following objective function

$$\begin{aligned} \min_{w \geq 0, h} \quad & -\log\det(\mathcal{L}w + J) + \operatorname{Tr}(\mathcal{L}wS_h) + \sum_i r_\lambda(w_i) \\ \text{s.t.} \quad & h^\intercal \mathbf{1} = \hat{g} \\ & h_i \in \{0, 1\} \text{ for all } i = 1, \ldots, n \\ & \|\mathcal{L}w\|_1 \leq R \end{aligned} \tag{3}$$

where $S_h = \frac{1}{\hat{g}} \sum_i h_i X^{(i)} X^{(i)^\intercal}$, h is a vector such that $h_i \in \{0, 1\}$ and $h^\intercal \mathbf{1} = \hat{g}$, and $\hat{g}$ is a hyperparameter that indicates our prior knowledge about how many "good" items we have in our dataset. Observe that the constraint $\|\mathcal{L}w\|_1 \leq R$ is rather mild in the sense that we can pick a large enough R so that $w^\star$ is feasible (Loh & Wainwright, 2013).

The objective function (3) is biconvex. We can therefore alternate between w and h by fixing h and fitting w, and vice versa. Provided a fixed h, we optimize for w using the Majorization-Minimization (MM) framework (Sun et al., 2017) as proposed by Ying et al. (2020). More specifically, we first define a surrogate function $f_k(w) = f(w|w^{(k)})$ that approximates the objective function of Equation (3) at the point $w^{(k)}$. Let $z_i^{(k)} = r'_\lambda(w_i^{(k)})$ be the derivative of r_λ at $w_i^{(k)}$. Then the surrogate function can be obtained by lin-

Algorithm 1 Trimmed LCGGM Learning

Input: $w^{(0)}, T, \hat{g}$
Initialize $h^{(0)}$
$S_{h^{(0)}} \leftarrow \frac{1}{\hat{g}} \sum_i h_i^{(0)} X^{(i)} X^{(i)^\intercal}$
repeat
  $z^{(k)} \leftarrow r'_\lambda(w_i^{(k)})$ for all $i = 1, \dots, p(p-1)/2$
  $\nabla f(w^{(k)}) \leftarrow -\mathcal{L}^*(\mathcal{L}w^{(k)} + J)^{-1} + \mathcal{L}^* S_{h^{(k)}} + z^{(k)}$
  Choose $\eta^{(k)}$ with line search.
  $w^{(k+1)} \leftarrow \left[w^{(k)} - \eta^{(k)} \nabla f(w^{(k)})\right]_+$
  if k mod $T = 0$ **then**
    Update $h^{(k)}$
    $S_{h^{(k)}} \leftarrow \frac{1}{\hat{g}} \sum_i h_i^{(k)} X^{(i)} X^{(i)^\intercal}$
  else
    $h^{(k)} \leftarrow h^{(k-1)}$ $\{h^{(k)}$ remains the same$\}$
    $S_{h^{(k)}} \leftarrow S_{h^{(k-1)}}$ $\{S_{h^{(k)}}$ remains the same$\}$
  end if
until Stopping criterion not satisfied

earizing the regularization function $\sum_i r_\lambda(w_i)$ as follows:

$$f_k(w) = -\log\det(\mathcal{L}w + J) + \mathrm{Tr}(\mathcal{L}w S_h) + \sum_i z_i^{(k)} w_i.$$

We then update $w^{(k)}$ by

$$w^{(k+1)} = \left[w^{(k)} - \eta^{(k)} \nabla f_k(w^{(k)})\right]_+$$

where $[a]_+ = \max(0, a)$, and where $\eta^{(k)}$ is the step size at the k^{th} iteration that can be found with line search. Updating h when w is fixed amounts to setting $h_i = 1$ if $X^{(i)}$ is among the $\hat{g}$ elements with the highest log-likelihood and $h_i = 0$ otherwise. In practice we can update h every T iterations until convergence, where T is a user input hyperparameter. The full procedure is described in Algorithm 1.

Choice of regularization function. In this work we restrict our analysis to two nonconvex type of penalties, namely MCP (Zhang, 2010) and SCAD (Fan & Li, 2001). These functions are defined as

$$r'_{\lambda;\mathrm{MCP}}(t) = \begin{cases} \lambda - \frac{t}{\beta} & \text{if } 0 \le t \le \beta\lambda \\ 0 & \text{if } t \ge \beta\lambda \end{cases}$$

and

$$r'_{\lambda;\mathrm{SCAD}}(t) = \begin{cases} \lambda & \text{if } 0 \le t \le \lambda \\ \frac{(\beta\lambda - t)}{\beta - 1} & \text{if } \lambda \le t \le \beta\lambda \\ 0 & \text{if } t \ge \beta\lambda \end{cases}$$

However our analysis can be extended to other penalization functions as long as they fulfill a set of conditions related to their unbiasedness and sparsity promoting nature (Ying et al., 2020; Loh & Wainwright, 2013; 2017).

3.2. Theoretical Analysis

Optimization of Equation (3) yields local optima due to the non-convexity of the objective function. In what follows we will prove that arbitrary local optima are guaranteed to be consistent under the Frobenius norm, provided some assumptions on the data and local optima are fulfilled. We first present in detail these assumptions.

The following lemma is often refered to as *restricted strong convexity* and it is crucial for deriving the error bound presented in this section.

Lemma 3.1. *(Ying et al. (2020)) Let w and $\hat{w}$ be such that $w, \hat{w} \ge 0$, $\mathcal{L}w + J$ and $\mathcal{L}\hat{w} + J$ are positive definite, and $\|\mathcal{L}w - \mathcal{L}\hat{w}\|_F \le r$. Then the following holds,*

$$\langle -\mathcal{L}^*(\mathcal{L}w + J)^{-1} + \mathcal{L}^*(\mathcal{L}\hat{w} + J)^{-1}, w - \hat{w} \rangle \ge (\|\mathcal{L}w^*\|_2 + r)^{-2} \|\mathcal{L}w - \mathcal{L}\hat{w}\|_F^2$$

Let $\kappa = (\|\mathcal{L}w^*\|_2 + 1)^{-2}$. We further assume the following for the coefficient λ.

Assumption 3.2. The following holds for the regularization coefficient λ,

$$4\max\{\|\mathcal{L}^*(\mathcal{L}w^* + J)^{-1} - S_{h^*}\|_\infty, \tau_1\} \le \lambda \le \left[\kappa - \frac{\tau_2 + \lambda k}{\sqrt{2}}\right] \frac{1}{R}.$$

The following lemma allows us to use Lemma 3.1 for arbitrary local optima of Equation (3).

Lemma 3.3. *Let $(\widetilde{w}, \widetilde{h})$ be any local optimum of Equation (3). Then, under Assumption 3.2 we have that*

$$\|\mathcal{L}\widetilde{w} - \mathcal{L}w^\star\|_F \le 1$$

The proof can be found in Appendix C.

We need to make further assumptions about local optima and the dataset in order to handle scenarios where the data is contaminated with outliers. The following condition is known as *mutual incoherence condition*:

Assumption 3.4. Let $(\widetilde{w}, \widetilde{h})$ be any local optimum of Equation (3). Then,

$$\|\mathrm{Tr}\{(S_{\widetilde{h}} - S_{h^\star})(\mathcal{L}\widetilde{w} - \mathcal{L}w^\star)\}\| \le \tau_1 \|\widetilde{w} - w^\star\|_1 + \tau_2 \|\widetilde{w} - w^\star\|_2.$$

We finally assume a bound on the outliers and on the number of outliers with respect to the number of "good" samples.

Assumption 3.5. Let $X^B \in \mathbb{R}^{b \times p}$ be the design matrix formed by the outliers. We assume $\|X^B\|_2^2 \le f(X^B)\sqrt{g \log p}$ for some function f (Yang & Lozano, 2015).

Assumption 3.6. The ratio between the amount of outliers and "good" samples is such that $\frac{b}{g} \le \frac{1}{\sqrt{n}}$.

Lemma 3.7. *(Nasrabadi et al., 2011; Yang & Lozano, 2015) If Assumption 3.5 holds then Assumption 3.4 holds with*

$$\tau_1 = f(X^B)\sqrt{\frac{\log p}{g}} \quad \text{and} \quad \tau_2 = f(X^B)\sqrt{\frac{b\log p}{g}}$$

We can now establish the consistency of *arbitrary* local optima found with Algorithm 1 under the Frobenius norm.

Theorem 3.8. *Let $(\widetilde{w}, \widetilde{h})$ be any local optimum of the objective in Equation (3). Let $\lambda = C\sqrt{\frac{\log p}{n}}$ where $C = 4\max\left\{2\sqrt{\alpha}\gamma + \frac{\gamma}{2\sqrt{2\log p}}, f(X^B)\right\}$, $\gamma = 2\sqrt{2}\|\mathcal{L}^*(\mathcal{L}w^\star + J)^{-1}\|_\infty$, and $D = \sqrt{p(p-1)/2}$. Then, under assumptions 3.2 - 3.6, if*

$$n \geq \left[\frac{f(X^B) + 3RC}{\kappa}\right]^2 \log p$$

it holds that

$$\|\mathcal{L}\widetilde{w} - \mathcal{L}w^\star\|_F \leq \frac{1}{\sqrt{2}\kappa}\left[\frac{3}{2}C\sqrt{\frac{\log p}{n}}D + \sqrt{\frac{b\log p}{g}}f(X^B)\right]$$

with probability at least $1 - 1/p^{\alpha-2}$ for some $\alpha > 2$.

The proof can be found in Appendix B.

4. Experiments

We empirically demonstrate that our method can effectively cope with the presence of outliers by conducting experiments on simulated data. Results on real world data can be found in Appendix D.

We considered two different simulated settings where the training sets were contaminated with items sampled from a distribution with a different graph structure. That is, we generated data from a mixture of Gaussians as

$$X \sim p_0\mathcal{N}(0, (\Theta^\star)^{-1}) + (1-p_0)\mathcal{N}(0, (\Theta_{Corr})^{-1})$$

where $\Theta^\star$ is the true parameter and p_0 is the probability of sampling a "good" element.

The scenarios we considered are:

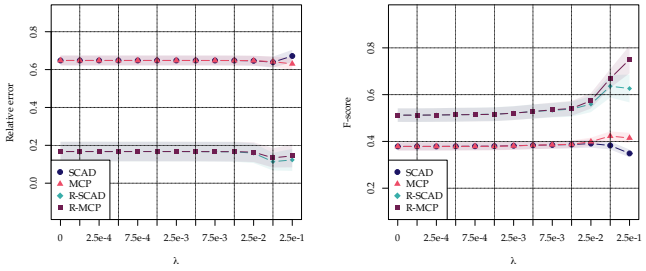
- Sc1: The graph structures for $\Theta^\star$ and Θ_{Corr} are Barabasi-Albert graphs of degre 1 generated with different seeds. For this scenario we set $p = 50$, $n = 250$, and $p_0 = 0.8$. The weights of the edges were uniformely sampled from $\mathcal{U}[0.25, 2]$.
- Sc2: $\Theta^\star$ is a graph with three hubs sampled from a stochastic block model with intra-cluster probability 0.05 and inter-cluster probability 0.003. Θ_{Corr} is a star graph. We fixed $p = 21$, $n = 105$, and $p_0 = 0.8$. The weights of the edges were uniformely sampled from $\mathcal{U}[1, 2.5]$.

We compare our methods R-SCAD and R-MCP to their non-robust counterparts, SCAD and MCP. For R-SCAD and R-MCP we tunned the $\hat{g}$ hyperparameter from a set of values such that $\hat{g}$ represents the 80%, 85% and 90% of the data. We fixed the update interval hyperparameter as $T = 50$. For comparison we use the Relative Error (RE) between the estimated parameter $\hat{\Theta}$ and the ground truth $\Theta^\star$, and the F-score defined as

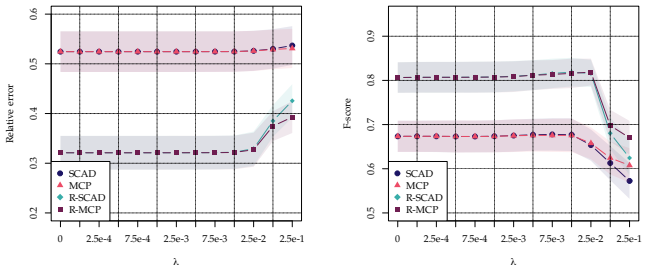
$$\text{RE} = \frac{\|\hat{\Theta} - \Theta^\star\|_F}{\|\Theta^\star\|_F} \quad \text{and} \quad \text{FS} = \frac{\text{tp}}{\text{tp} + 0.5(\text{fp} + \text{fn})}$$

respectively.

Figure 1 summarizes the mean and standard deviation of RE and FS in both scenarios for different values of λ over 100 simulations. We observe that R-SCAD and R-MCP are systematically better that their non-robust counterparts in terms of relative error and F-score.



(a) RE for Sc1 (b) FS for Sc1



(c) RE for Sc2 (d) FS for Sc2

Figure 1. Mean relative error and F-score for scenarios Sc1 and Sc2, for different values of λ.

References

Banerjee, O., d'Aspremont, A., and Ghaoui, L. E. Sparse covariance selection via robust maximum likelihood estimation. *ArXiv*, 2005.

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A. D., and Vandergheynst, P. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 2017.

Cai, T. T., Liu, W., and Zhou, H. H. Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *The Annals of Statistics*, 2016.

Dong, X., Thanou, D., Frossard, P., and Vandergheynst, P. Learning laplacian matrix in smooth graph signal representations. *IEEE Transactions on Signal Processing*, 2016.

Dong, X., Thanou, D., Toni, L., Bronstein, M. M., and Frossard, P. Graph signal processing for machine learning: A review and new perspectives. *IEEE Signal Processing Magazine*, 2020.

Egilmez, H. E., Pavez, E., and Ortega, A. Graph learning from data under laplacian and structural constraints. *IEEE Journal of Selected Topics in Signal Processing*, 2017.

Fan, J. and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 2001.

Friedman, J., Hastie, T., and Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 2008.

Kalofolias, V. How to learn a graph from smooth signals. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research.

Koyakumaru, T., Yukawa, M., Pavez, E., and Ortega, A. Learning sparse graph with minimax concave penalty under gaussian markov random fields. *ArXiv*, 2021.

Kumar, S., Ying, J., de Miranda Cardoso, J. V., and Palomar, D. Structured graph learning via laplacian spectral constraints. In *Advances in Neural Information Processing Systems*, 2019.

Lake, B. M. and Tenenbaum, J. B. Discovering structure by learning sparse graphs. 2010.

Loh, P.-L. and Wainwright, M. J. Regularized m-estimators with nonconvexity: statistical and algorithmic theory for local optima. In *Journal of Machine Learning Research*, 2013.

Loh, P.-L. and Wainwright, M. J. Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics*, 2017.

Nasrabadi, N., Tran, T., and Nguyen, N. Robust lasso with missing and grossly corrupted observations. In *Advances in Neural Information Processing Systems*, 2011.

Ortega, A., Frossard, P., Kovačević, J., Moura, J. M. F., and Vandergheynst, P. Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, 2018.

Pavez, E. Laplacian constrained precision matrix estimation: Existence and high dimensional consistency. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022.

Shuman, D. I., Narang, S. K., Frossard, P., Ortega, A., and Vandergheynst, P. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 2013.

Sun, Y., Babu, P., and Palomar, D. P. Majorization-minimization algorithms in signal processing, communications, and machine learning. *IEEE Transactions on Signal Processing*, 2017.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science*, 2000.

von Luxburg, U. A tutorial on spectral clustering. *Statistics and Computing*, 2007.

Yang, E. and Lozano, A. C. Robust gaussian graphical modeling with the trimmed graphical lasso. In *Advances in Neural Information Processing Systems*, 2015.

Ying, J., de Miranda Cardoso, J. V., and Palomar, D. Nonconvex sparse graph learning under laplacian constrained graphical model. In *Advances in Neural Information Processing Systems*, 2020.

Ying, J., Vinícius de Miranda Cardoso, J., and Palomar, D. Minimax estimation of laplacian constrained precision matrices. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, 2021.

Zhang, C.-H. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38, 2010.

A. Useful lemma

Lemma A.1. *(Ying et al. (2020)) Let $\lambda = \sqrt{4\alpha \log p/n}\gamma$, where γ is defined as in Theorem 3.8. Let $n \geq 8\alpha \log p$ for some $\alpha > 2$. Then*

$$P[\|\mathcal{L}^*(\mathcal{L}w^\star + J)^{-1} - \mathcal{L}^* S\|_\infty \leq \lambda/2] \geq 1 - \frac{1}{p^{\alpha-2}}$$

B. Proof of Theorem 1

Proof. Let $(\widetilde{w}, \widetilde{h})$ be any local optimum of Equation (3). Lemma 3.3 tells us that $\|\mathcal{L}\widetilde{w} - \mathcal{L}w^\star\|_F \leq 1$. We can then apply Lemma 3.1 with $r = 1$. Let $\kappa = (\|\mathcal{L}w^*\|_2 + 1)^{-2}$. Then,

$$\|\mathcal{L}\widetilde{w} - \mathcal{L}w^\star\|_F^2 \leq \frac{1}{\kappa}\langle -\mathcal{L}^*(\mathcal{L}\widetilde{w} + J)^{-1} + \mathcal{L}^*(\mathcal{L}w^\star + J)^{-1}, \widetilde{w} - w^\star\rangle \tag{4}$$

On the other hand we have

$$0 \leq \langle \mathcal{L}^*(\mathcal{L}\widetilde{w} + J)^{-1} - \mathcal{L}^* S_{\widetilde{h}}, \widetilde{w} - w^\star\rangle - \langle \widetilde{z}, \widetilde{w} - w^\star\rangle \tag{5}$$

which follows from the fact that $(\widetilde{w}, \widetilde{h})$ is a local optimum. Combining equations (4) and (5) we obtain

$$\begin{aligned}\|\mathcal{L}\widetilde{w} - \mathcal{L}w^\star\|_F^2 &\leq \frac{1}{\kappa}\langle \mathcal{L}^*(\mathcal{L}w^\star + J)^{-1} - \mathcal{L}^* S_{\widetilde{h}}, \widetilde{w} - w^\star\rangle - \langle \widetilde{z}, \widetilde{w} - w^\star\rangle \\ &= \frac{1}{\kappa}\underbrace{\langle \mathcal{L}^*(\mathcal{L}w^\star + J)^{-1} - \mathcal{L}^* S_{h^\star}, \widetilde{w} - w^\star\rangle}_{(i)} + \underbrace{\langle \mathcal{L}^* S_{h^\star} - \mathcal{L}^* S_{\widetilde{h}}, \widetilde{w} - w^\star\rangle}_{(ii)} - \underbrace{\langle \widetilde{z}, \widetilde{w} - w^\star\rangle}_{(iii)}\end{aligned}$$

We now proceed to bound all three terms above.

To bound (i) we apply Hölder's inequality and Assumption 3.2:

$$\begin{aligned}\langle \mathcal{L}^*(\mathcal{L}w^\star + J)^{-1} - \mathcal{L}^* S_{h^\star}, \widetilde{w} - w^\star\rangle &\leq \|\mathcal{L}^*(\mathcal{L}w^\star + J)^{-1} - \mathcal{L}^* S_{h^\star}\|_\infty \|\widetilde{w} - w^\star\|_1 \\ &\leq \frac{1}{4}\lambda\|\widetilde{w} - w^\star\|_1\end{aligned}$$

To bound (ii), note that by Assumption 3.4 we have

$$\begin{aligned}\langle \mathcal{L}^* S_{h^\star} - \mathcal{L}^* S_{\widetilde{h}}, \widetilde{w} - w^\star\rangle &\leq \tau_1\|\widetilde{w} - w^\star\|_1 + \tau_2\|\widetilde{w} - w^\star\|_2 \\ &\leq \frac{1}{4}\lambda\|\widetilde{w} - w^\star\|_1 + \tau_2\|\widetilde{w} - w^\star\|_2\end{aligned}$$

Finally, we use Cauchy-Schwarz to bound (iii) as follows:

$$\begin{aligned}\langle \widetilde{z}, \widetilde{w} - w^\star\rangle &\geq -\|\widetilde{z}\|_2\|\widetilde{w} - w^\star\|_2 \\ -\langle \widetilde{z}, \widetilde{w} - w^\star\rangle &\leq \|\widetilde{z}\|_2\|\widetilde{w} - w^\star\|_2 \\ &\leq \sqrt{p(p-1)/2}\|\widetilde{z}\|_\infty\|\widetilde{w} - w^\star\|_2 \\ &\leq \sqrt{p(p-1)/2}\lambda\|\widetilde{w} - w^\star\|_2\end{aligned}$$

Let $D = \sqrt{p(p-1)/2}$. Putting all together we obtain

$$\begin{aligned}\|\mathcal{L}\widetilde{w} - \mathcal{L}w^\star\|_F^2 &\leq \frac{1}{\kappa}\left[\frac{\lambda}{2}\|\widetilde{w} - w^\star\|_1 + (\tau_2 + D\lambda)\|\widetilde{w} - w^\star\|_2\right] \\ &\leq \frac{1}{\kappa}\left[\left(\frac{3}{2}D\lambda + \tau_2\right)\|\widetilde{w} - w^\star\|_2\right] \\ &\leq \frac{1}{\kappa}\left(\frac{3}{2}D\lambda + \tau_2\right)\frac{1}{\sqrt{2}}\|\mathcal{L}\widetilde{w} - \mathcal{L}w^\star\|_F\end{aligned}$$

where the last inequality follows from the fact that $\sqrt{2}\|\widetilde{w} - w^\star\|_2 \leq \|\mathcal{L}\widetilde{w} - \mathcal{L}w^\star\|_F$. It follows that

$$\|\mathcal{L}\widetilde{w} - \mathcal{L}w^\star\|_F \leq \frac{1}{\kappa\sqrt{2}}\left(\frac{3}{2}D\lambda + \tau_2\right)$$

□

C. Proof of Lemma 1

Proof. Assume $\|\mathcal{L}\widetilde{w} - \mathcal{L}w^\star\|_F \geq 1$. Let $w_t = w^\star + t(\widetilde{w} - w^\star)$. Take $t = 1/\|\mathcal{L}\widetilde{w} - \mathcal{L}w^\star\|_F$. Then $\|t(\mathcal{L}\widetilde{w} - \mathcal{L}w^\star)\|_F = 1$ and hence we can apply Lemma 3.1:

$$\left\langle -\mathcal{L}^*(\mathcal{L}w_t + J)^{-1} + \mathcal{L}^*(\mathcal{L}w^\star + J)^{-1}, w_t - w^\star \right\rangle \geq (\|\mathcal{L}w^\star\|_2 + 1)^{-2}\|\mathcal{L}w_t - \mathcal{L}w^\star\|_F^2,$$

then

$$\begin{aligned} t\left\langle -\mathcal{L}^*(\mathcal{L}\widetilde{w} + J)^{-1} + \mathcal{L}^*(\mathcal{L}w^\star + J)^{-1}, w_t - w^\star \right\rangle &\geq (\|\mathcal{L}w^\star\|_2 + 1)^{-2}\|\mathcal{L}w_t - \mathcal{L}w^\star\|_F \\ &= (\|\mathcal{L}w^\star\|_2 + 1)^{-2}t\|\mathcal{L}\widetilde{w} - \mathcal{L}w^\star\|_F^2, \end{aligned}$$

so we have

$$\left\langle -\mathcal{L}^*(\mathcal{L}\widetilde{w} + J)^{-1} + \mathcal{L}^*(\mathcal{L}w^\star + J)^{-1}, w_t - w^\star \right\rangle \geq (\|\mathcal{L}w^\star\|_2 + 1)^{-2}\|\mathcal{L}\widetilde{w} - \mathcal{L}w^\star\|_F^2. \tag{6}$$

Since $(\widetilde{w}, \widetilde{h})$ is a local optimum we have,

$$0 \leq \left\langle \mathcal{L}^*(\mathcal{L}\widetilde{w} + J)^{-1} - \mathcal{L}^*S_{\widetilde{h}}, \widetilde{w} - w^\star \right\rangle - \left\langle \widetilde{z}, \widetilde{w} - w^\star \right\rangle. \tag{7}$$

Combining (6) and (7) we obtain

$$(\|\mathcal{L}w^\star\|_2 + 1)^{-2}t\|\mathcal{L}\widetilde{w} - \mathcal{L}w^\star\|_F^2 \leq \left\langle \mathcal{L}^*(\mathcal{L}w^\star + J)^{-1} - \mathcal{L}^*S_{\widetilde{h}} - \widetilde{z}, \widetilde{w} - w^\star \right\rangle.$$

After adding and subtracting $\mathcal{L}^*S_{h^\star}$ and rearranging, the right hand side becomes

$$\underbrace{\left\langle \mathcal{L}^*(\mathcal{L}w^\star + J)^{-1} - \mathcal{L}^*S_{h^\star}, \widetilde{w} - w^\star \right\rangle}_{(i)} + \underbrace{\left\langle \mathcal{L}^*S_{h^\star} - \mathcal{L}^*S_{\widetilde{h}}, \widetilde{w} - w^\star \right\rangle}_{(ii)} - \underbrace{\left\langle \widetilde{z}, \widetilde{w} - w^\star \right\rangle}_{(iii)}$$

To bound (i) we apply Hölder's inequality and Assumption 3.2:

$$\begin{aligned} \left\langle \mathcal{L}^*(\mathcal{L}w^\star + J)^{-1} - \mathcal{L}^*S_{h^\star}, \widetilde{w} - w^\star \right\rangle &\leq \|\mathcal{L}^*(\mathcal{L}w^\star + J)^{-1} - \mathcal{L}^*S_{h^\star}\|_\infty \|\widetilde{w} - w^\star\|_1 \\ &\leq \frac{1}{4}\lambda\|\widetilde{w} - w^\star\|_1 \end{aligned}$$

To bound (ii), we use Assumption 3.4:

$$\begin{aligned} \left\langle \mathcal{L}^*S_{h^\star} - \mathcal{L}^*S_{\widetilde{h}}, \widetilde{w} - w^\star \right\rangle &\leq \tau_1\|\widetilde{w} - w^\star\|_1 + \tau_2\|\widetilde{w} - w^\star\|_2 \\ &\leq \frac{1}{4}\lambda\|\widetilde{w} - w^\star\|_1 + \tau_2\|\widetilde{w} - w^\star\|_2 \end{aligned}$$

Finally, we use Cauchy-Schwarz to bound (iii) as follows:

$$\begin{aligned} \left\langle \widetilde{z}, \widetilde{w} - w^\star \right\rangle &\geq -\|\widetilde{z}\|_2\|\widetilde{w} - w^\star\|_2 \\ -\left\langle \widetilde{z}, \widetilde{w} - w^\star \right\rangle &\leq \|\widetilde{z}\|_2\|\widetilde{w} - w^\star\|_2 \\ &\leq \sqrt{p(p-1)/2}\|\widetilde{z}\|_\infty\|\widetilde{w} - w^\star\|_2 \\ &\leq \sqrt{p(p-1)/2}\lambda\|\widetilde{w} - w^\star\|_2 \end{aligned}$$

Putting all terms together and using our assumption on λ yields the result. □

Table 1.

λ	SCAD	R-SCAD
0.0001	18.73	**16.41**
0.001	18.62	**11.58**
0.01	**18.03**	21.35
0.1	**16.75**	16.84
0.5	**10.18**	14.40

Table 2.

λ	MCP	R-MCP
0.0001	18.74	**17.93**
0.001	18.70	**11.69**
0.01	**18.15**	18.67
0.1	17.44	**16.49**
0.5	11.90	**7.67**

Table 3. Percentage of "bad" edges in SCAD and R-SCAD (left), and MCP and R-MCP (right) for different values of λ.

D. Results on Real-world Data

We also carried experiments on the `2019-nCoV` dataset available in the R package `nCov2019`. The dataset collects information about 98 Chinese patients who were affected by the Covid infection. Patients are classified into those who survived the disease and those who did not.

We use the same experimental setup as in Ying et al. (2020), that is, we transformed categorical features into one-hot encodings, which results in a design matrix $X \in \mathbb{R}^{32\times 98}$. Our goal is to learn a graph from the data where the nodes represent patients and edges encode a notion of similarity between patients. A desirable property for a graph is for it to connect nodes that belong to the same class membership, "survived" or "did not survive". Note that this behavior encodes a notion of smoothness of the graph with respect to the survival status of the patients.

To compare R-SCAD and R-MCP to their non-robust counterparts, SCAD and MCP, we estimated graphs for each method for different values of λ and collected the percentage of "bad" edges with respect to the total number of recovered edges. A "bad" edge in this context is an edge that connects two nodes belonging to different class memberships. For both robust methods we fixed the hyperparameter $\hat{g} = 29$, and $T = 50$. Table 3 shows that graphs estimated with the robust methods are in general better at capturing the different node clusters, in contrast to non-robust methods. R-MCP generally outperforms MCP by a considerable margin and it is on a par with MCP for $\lambda = 0.01$. Although R-SCAD lags behind SCAD for two values of λ, it outperforms SCAD by a large margin for smaller values of λ, and it is on a par with SCAD for $\lambda = 0.1$. The fact that robust methods better capture node clusters for many different values of λ demonstrates that this "grouping effect" does not depend on the sparsity of the estimated graph.