Computational Statistics (2023) 38:1337–1355
https://doi.org/10.1007/s00180-022-01265-w

ORIGINAL PAPER

Sparse precision matrix estimation with missing observations

Ning Zhang[1] · **Jin Yang**[1]

Received: 2 December 2021 / Accepted: 13 July 2022 / Published online: 26 July 2022

Abstract
Sparse Gaussian graphical models have been extensively applied to detect the conditional independence structures from fully observed data. However, datasets with missing observations are quite common in many practical fields. In this paper, we propose a robust Gaussian graphical model with the covariance matrix being estimated from the partially observed data. We prove that the inverse of the Karush–Kuhn–Tucker mapping associated with the proposed model satisfies the calmness condition automatically. We also apply a linearly convergent alternating direction method of multipliers to find the solution to the proposed model. The numerical performance is evaluated on both the synthetic data and real data sets.

Keywords Missing data · Inverse probability weighting · Gaussian graphical model · ADMM

1 Introduction

Let $\{\xi_i \in \mathbb{R}^n,\ i = 1, \ldots, m\}$ be a set of samples independently drawn from a Gaussian distribution $\mathcal{N}(\mu, \Sigma)$. It is well known that the elements of a precision matrix (i.e., the inverse of covariance matrix Σ^{-1}) could be applied to characterize the conditional independence between two variables. This is based on the fact proved in Dempster (1972) that the i-th and j-th components of random variable $\xi \sim \mathcal{N}(\mu, \Sigma)$ are conditionally independent if and only if $[\Sigma^{-1}]_{ij} = 0$. However, the true covariance matrix Σ is hardly known in practice, but usually can be estimated from the observations.

The research of this author was supported by the National Natural Science Foundation of China (11901083, 12171153), Guangdong Basic and Applied Basic Research Foundation (2022A1515010088).

✉ Ning Zhang
zhangning@dgut.edu.cn

[1] School of Computer Science and Technology, Dongguan University of Technology, Dongguan 523808, China

When the samples are fully observed, let $\hat{\mu} := \frac{1}{m}\sum_{i=1}^{m}\xi_i$, the sample covariance matrix $\hat{s} = \frac{1}{m}\sum_{i=1}^{m}(\xi_i - \hat{\mu})(\xi_i - \hat{\mu})^T$ is a widely used estimator. If the sample size m is larger than the dimension n, the sample covariance matrix $\hat{s}$ is in general positive definite, but the elements of precision matrix estimator $\hat{s}^{-1}$ are typically nonzero (Yuan and Lin 2007). Moreover, contemporary datasets are often high dimensional ($n \gtrsim m$). In this setting, the sample covariance matrix is rank deficient. Therefore, the sparse Gaussian graphical model (sGGM) was introduced (Yuan and Lin 2007) :

$$\min_{x>0} -\log\det x + \langle \hat{s}, x\rangle + \lambda\|x\|_1, \tag{1.1}$$

where $\lambda > 0$, $\hat{s}$ is the sample covariance matrix estimated by the fully observed samples, and $\|x\|_1 = \sum_{i,j=1}^{n}|x_{ij}|$. The sparse Gaussian graphical model and its variants have been applied in many fields, such as climate networks (Zerenner et al. 2014), biological networks (Wang et al. 2016; Zhang et al. 2018), and traffic management (Sun et al. 2012). A large amount of literature is devoted to the algorithm design and analysis: interior point methods (Lu and Toh 2010; Yuan and Lin 2007), block coordinate descent method (Friedman et al. 2008), alternating direction method of multipliers (Yuan 2012), Newton-type methods (Hsieh et al. 2014; Wang et al. 2010), to name just a few.

Dataset with missing values is a ubiquitous phenomenon in the practical world (Lounici 2014). The covariance matrix is an essential part of many applications and algorithms (Pavez and Ortega 2021). However, only a few literature study the covariance estimation from the data with missing values. Inverse probability weighting (IPW) approaches (Seaman and White 2013) are commonly applied to correct the bias of the covariance estimation from the partially observed data. The estimator obtained by the IPW approach is usually named as IPW estimator. Under the assumption that the data are missing completely at random, Kolar and Xing (2012), and Lounici (2014) introduced different IPW estimators instead of the sample covariance matrix. Recently, under more general assumptions about missingness, new IPW estimators are proposed by Park and Lim (2019), Park et al. (2020), Pavez and Ortega (2021). Moreover, the precision matrix estimation based on the IPW estimator has also been studied in Fan et al. (2019); Kolar and Xing (2012).

It has been proved that problem (1.1) has a unique solution under the assumption that the covariance matrix estimation is positive semidefinite (Hsieh et al. 2014). However, the IPW estimators are usually non-positive semidefinite. When the non-positive semidefinite IPW estimator is taken as a surrogate of the sample covariance matrix, the objective function in problem (1.1) could be unbounded from below. Therefore, problem (1.1) may fail to yield a precision matrix estimator when $\hat{s}$ is non-positive semidefinite. Moreover, the estimation of the covariance matrix from data with missing values is still challenging work.

In this paper, we consider the following Gaussian graphical model:

$$\min_{x\in\mathcal{C}} -\log\det x + \langle \hat{s}, x\rangle + \lambda\|x\|_{1,\mathrm{off}}, \tag{1.2}$$

where $\hat{s} \in \mathbb{S}^n$ is a given IPW estimator (not necessarily positive semidefinite), $\|x\|_{1,\text{off}} = \sum_{i \neq j} |x_{ij}|$ and $\mathcal{C} := \{x \in \mathbb{S}^n : \|x\| \leq \alpha\}$ with $\|x\| := (\sum_{i,j}^n x_{ij}^2)^{1/2}$ and $\alpha > 0$. The main motivations to study problem (1.2) are listed below:

- *Lower boundedness* the set $\mathcal{C}$, which is named as side constraint in Loh and Wainwright (2015), is used to guarantee the existence of a unique global optimal solution.
- *Robustness* the proposed model (1.2) can be used to hedge against the risk raised by the uncertainty of the covariance matrix estimation. Specifically, there exists a positive scalar μ such that problem (1.2) can be equivalently rewritten as the following penalized problem:

$$\min_{x \succ 0} -\log\det x + \langle \hat{s}, x\rangle + \lambda \|x\|_{1,\text{off}} + \mu \|x\|. \tag{1.3}$$

  Furthermore, by using the identity $\mu\|x\| = \max_{\|s-\hat{s}\| \leq \mu} \langle s - \hat{s}, x\rangle$, problem (1.3) can be equivalently reformulated into the following robust counterpart of sparse Gaussian graphical model,

$$\min_{x \succ 0} \max_{s \in \mathcal{U}} -\log\det x + \langle s, x\rangle + \lambda \|x\|_{1,\text{off}},$$

  where $\mathcal{U}$ is the uncertainty set defined as

$$\mathcal{U} := \{s \in \mathbb{S}^n : \|s - \hat{s}\| \leq \mu\}, \ \mu > 0.$$

  Therefore, we refer to the problem (1.2) as a robust Gaussian graphical model (rGGM) throughout this paper.

In addition to the lower boundedness and robustness, another motivation to consider rGGM is to pave the way for solving nonconvex regularized Gaussian graphical model [see e.g., Fan et al. (2019), Loh and Wainwright (2015)]. It has been commonly accepted that nonconvex regularizers have better performance than that of convex regularizers (Fan et al. 2016). Besides, the widely used smoothly clipped absolute deviation [SCAD, Fan and Li (2001)] and minimax concave penalty [MCP, Zhang (2010)] regularizers can be reformulated as the difference of ℓ_1 norm and some smooth convex functions (Ahn et al. 2017; Tang et al. 2020). Note that the lower boundedness assumption usually plays an essential role in the theoretical analysis of various numerical algorithms. Therefore, a systematic study on the convex regularized Gaussian graphical model with side constraints may provide a theoretical foundation for designing more efficient numerical algorithms for solving the nonconvex regularized Gaussian graphical models.

Note that the proposed Gaussian graphical model is a composite convex optimization problem:

$$\min_{x \succ 0} -\log\det x + \langle \hat{s}, x\rangle + p(x), \quad p(x) := \lambda \|x\|_{1,\text{off}} + \mathbb{I}_{\mathcal{C}}(x). \tag{1.4}$$

The alternating direction method of multipliers (ADMM) has been extensively used for solving the structured composite convex optimization problem [see e.g. Fan et al. (2019), Yuan (2012), Yuan et al. (2020)]. Under some calmness conditions,

(Han et al. 2018) show that the globally convergent semiproximal ADMM can achieve a linear convergence rate. A systematical study on the linear convergence of various ADMM-type algorithms has been given in Yuan et al. (2020). However, due to the side constraint $\mathcal{C}$, the linear convergence rate of ADMM-type algorithms for solving problem (1.2) can not be directly obtained from Yuan et al. (2020, Table 4). Therefore, we investigate the stability analysis of the solution mapping associated with the problem (1.2). This will facilitate the algorithm design and give the theoretical guarantee of various algorithms, such as the alternating direction method of multipliers and the proximal point algorithm.

The remaining parts of this paper are organized as follows. In Sect. 2, we present some definitions and preliminary results. In particular, we derive the specific expression of the proximal mapping associated with the convex regularizer p and the Lipschitz continuity of the KKT solution mapping associated with the problem (1.2). In Sect. 3, we propose the alternating direction method of multipliers, its implementation details, and convergence analysis for solving the problem (1.2). We evaluate the numerical performance of synthetic data and real data in Sect. 4 and conclude the paper in Sect. 5.

Here, we collect the frequently used notations of this paper. Let $\mathbb{S}^n$ ($\mathbb{S}^n_+, \mathbb{S}^n_{++}$) be the space of $n \times n$ real symmetric (positive semidefinite, positive definite) matrices. For a matrix $x \in \mathbb{S}^n$, we use $\|x\|_1$ and $\|x\|$ to denote its vector ℓ_1 and Frobenious norm, i.e., $\|x\|_{1,\text{off}} = \sum_{i\neq j} |x_{ij}|$ and $\|x\| = (\sum_{i,j} x_{ij}^2)^{1/2}$. Let $\mathbb{I}_C$ denote the indicator function of the convex set $C \subseteq \mathbb{S}^n$, that is $\mathbb{I}_C(x) = 0$ if $x \in C$, and $+\infty$ otherwise, and $\Pi_C(x)$ denote the Euclidean projection of x onto C.

2 Tools and definitions

In this section, we recall some definitions and present the results that will be used in the theoretical analysis and numerical implementation. Let $\mathbb{X}$, $\mathbb{Y}$ be two finite-dimensional real Hilbert spaces.

2.1 Proximal mapping

The proximal mapping associated with the regularizer p, a sum of two functions, plays a key role in the arithmetic design.

Definition 2.1 (Moreau 1965; Yosida 1964 *regularization*.) Let $f : \mathbb{X} \to \mathfrak{R} \cup \{+\infty\}$ be a proper, lower semicontinuous, convex function. The Moreau-Yosida regularization of f is given by

$$\Phi_f(x) := \min_{z\in\mathbb{X}} \left\{ f(z) + \frac{1}{2}\|z - x\|^2 \right\}, \ x \in \mathbb{X}. \tag{2.1}$$

The unique solution to problem (2.1) is called the proximal mapping associated with f which is defined as

$$\mathrm{Prox}_f(x) := \arg\min_{z\in\mathbb{X}} \left\{ f(z) + \frac{1}{2}\|z-x\|^2 \right\},\ x \in \mathbb{X}.$$

Lemma 2.1 *Let* $p : \mathbb{S}^n \to \mathbb{R} \cup \{+\infty\}$ *be defined by* $p(x) = \lambda\|x\|_{1,\mathrm{off}} + \mathbb{I}_{\mathcal{C}}(x)$, *then one has*

$$\mathrm{Prox}_p(x) = \Pi_{\mathcal{C}} \circ \mathrm{Prox}_{\lambda\|\cdot\|_{1,\mathrm{off}}}(x),\ \ \forall x \in \mathbb{S}^n.$$

Proof The proof is presented in Appendix 6.1. □

2.2 Lipschitz-like properties

The Lipschitz-like properties of solution mapping corresponding to convex optimization problems can be employed to establish the convergence rates of various algorithms. Firstly, we consider the following linearly perturbed formulation:

$$\begin{array}{ll} \min\limits_{x,y} & f(x) + \langle \hat{s}, x\rangle + p(y) - \langle (u,v),(x,y)\rangle \\ \text{s.t.} & x - y + w = 0, \end{array} \tag{2.2}$$

where $f(x) := -\log\det x$ and $p : \mathbb{S}^n \to \mathfrak{R} \cup \{+\infty\}$ is defined by (1.4). Define a mapping:

$$\mathcal{K}_{pert}((x,y,z);(u,v,w)) = \begin{pmatrix} x - \mathrm{Prox}_f(x+u-z-\hat{s}) \\ y - \mathrm{Prox}_p(y+v+z) \\ x-y-w \end{pmatrix}, \tag{2.3}$$

and a multifunction $\mathsf{Sol} : \mathbb{S}^n \times \mathbb{S}^n \times \mathbb{S}^n \rightrightarrows \mathbb{S}^n \times \mathbb{S}^n \times \mathbb{S}^n$:

$$\begin{aligned} \mathsf{Sol}(u,v,w) &:= \left\{ (x,y,z) : \mathcal{K}_{pert}((x,y,z);(u,v,w)) = 0 \right\} \\ &= \text{set of all}(x,y,z)\text{satisfying the KKT condition for problem (2.2).} \end{aligned} \tag{2.4}$$

Proposition 2.1 *The multifunction* Sol *defined by* (2.4) *is Lipschitz continuous near the origin, i.e., there exist a neighborhood* $\mathcal{N}$ *of the origin and a positive scalar* κ *such that* $\mathsf{Sol}(u,v,w) \neq \emptyset$ *for any* $(u,v,w) \in \mathcal{N}$ *and*

$$\|\mathsf{Sol}(u,v,w) - \mathsf{Sol}(u',v',w')\| \le \kappa\|(u,v,w) - (u',v',w')\|,\ \forall\,(u,v,w),(u',v',w') \in \mathcal{N}.$$

Proof The proof is given in Appendix 6.2. □

The proof of Proposition 2.1 is motivated by the routine of the proof in Zhang et al. (2020, Theorem 3.1). However, the regularizer p defined in (1.4) is no longer positive homogeneous, which is beyond the scope of the framework presented in Zhang et al. (2020). Therefore, based on Proposition 2.1, the proximal point dual Newton algorithm (Zhang et al. 2020) potentially can be applied to solve the problem (1.2).

To establish the linear convergence rate of the alternating direction method of multipliers for solving problem (1.2) based on the results presented in Han et al.

(2018), we should recall the definition of the calmness of a set-valued mapping $F : \mathbb{X} \rightrightarrows \mathbb{Y}$.

Definition 2.2 A set-valued mapping $F : \mathbb{X} \rightrightarrows \mathbb{Y}$ is said to be calm at $(\bar{x}, \bar{y}) \in \mathrm{gph} F$ with modulus $\kappa \geq 0$ if there exist neighborhoods $\mathcal{N}_{\bar{x}}$ of $\bar{x}$ and $\mathcal{N}_{\bar{y}}$ of $\bar{y}$ such that

$$F(x) \cap \mathcal{N}_{\bar{y}} \subseteq F(\bar{x}) + \kappa \|x - \bar{x}\| \mathbb{B}, \ \ \forall\, x \in \mathcal{N}_{\bar{x}},$$

where $\mathbb{B}$ denotes the closed unit ball in $\mathbb{Y}$.

The definition was first introduced as the pseudo-upper Lipschitz continuity in Ye and Ye (1997, Definition 2.8) and was coined as calmness in Rockafellar and Wets (1998). It is well known that $\mathcal{F}$ is calm at $(\bar{x}, \bar{y}) \in \mathrm{gph}\, \mathcal{F}$ if and only if its inverse set-valued mapping $\mathcal{F}^{-1}$ is metrically subregular [c.f. Dontchev and Rockafellar (2004, Definition 3.1)] at $(\bar{y}, \bar{x}) \in \mathrm{gph}\, \mathcal{F}^{-1}$.

3 ADMM algorithm

In this section, we present the Alternating Direction Method of Multipliers (ADMM) for solving the problem (1.2) and prove that the proposed method is globally linearly convergent. By introducing auxiliary variable, problem (1.2) can be reformulated as

$$\begin{aligned} &\min_{x,y} \ f(x) + \langle \hat{s}, x \rangle + p(y) \\ &\text{s.t.} \ \ x - y = 0. \end{aligned} \tag{3.1}$$

For a given positive scalar $\sigma > 0$, the augmented Lagrangian function associated with problem (3.1) is defined by

$$\mathbb{L}_\sigma(x, y, z) = f(x) + \langle \hat{s}, x \rangle + p(y) + \langle x - y, z \rangle + \frac{\sigma}{2} \|x - y\|^2.$$

The Karush–Kuhn–Tucker (KKT) condition of problem (3.1) can be described as follows:

$$\nabla f(x) + z + \hat{s} = 0, \ \ 0 \in \partial p(y) - z, \ \ \text{and} \ \ x - y = 0. \tag{3.2}$$

The ADMM for problem (3.1) can be characterized as follows:

Algorithm 1 Alternating Direction Method of Multipliers

Input $x^0 \in \mathbb{S}^n_{++}, y^0, z^0 \in \mathbb{S}^n$ and $\tau \in (0, (1 + \sqrt{5})/2)$. Iterate the following steps for $k = 0, 1, \ldots$:

Repeat

Step 1. Compute $y^{k+1} = \arg\min_y \mathbb{L}_\sigma(x^k, y, z^k)$.

Step 2. Compute $x^{k+1} = \arg\min_x \mathbb{L}_\sigma(x, y^{k+1}, z^k)$.

Step 3. Update $z^{k+1} = z^k + \tau\sigma(x^{k+1} - y^{k+1})$.

Until $(x^{k+1}, y^{k+1}, z^{k+1})$ satisfies the preset stopping conditions.

3.1 Convergence analysis

The ADMM algorithm has been widely applied in convex regularized Gaussian graphical model. Its globally convergent properties have been well-established. However, to the best of our knowledge, there is lack of systematic analysis on the convergence rate of ADMM for solving the rGGM. In this part, we show that ADMM for solving the problem (3.1) is globally and linearly convergent.

Define a KKT mapping $R : \mathbb{S}^n \times \mathbb{S}^n \times \mathbb{S}^n \to \mathbb{S}^n \times \mathbb{S}^n \times \mathbb{S}^n$:

$$R(x, y, z) = \begin{pmatrix} x - \underset{f}{\text{Prox}}\,(x - z - \hat{s}) \\ y - \underset{p}{\text{Prox}}(y + z) \\ x - y \end{pmatrix}. \tag{3.3}$$

Then, it holds that

$$R(x, y, z) = 0 \Longleftrightarrow (x, y, z)\text{satisfies the KKT condition(3.2)}.$$

That is, $R^{-1}(0)$ is the set of KKT points and $R^{-1}(0) = \mathsf{Sol}(0, 0, 0)$.

The following result plays a key role in establishing the linear convergence rate of ADMM. Its proof relies on several auxiliary results, we give the details in Appendix 6.2.

Proposition 3.1 *Let R be the KKT mapping defined by (3.3), then R^{-1} is calm at the origin for the KKT point $(\bar{x}, \bar{y}, \bar{z})$ of problem (3.1).*

Proof From Definition 2.2, it is sufficient to show that there exist neighborhoods $\mathcal{W}$ of $(\bar{x}, \bar{y}, \bar{z})$ and $\mathcal{V}$ of origin, a positive scalar κ_0 such that

$$R^{-1}(u, v, w) \cap \mathcal{W} \subseteq R^{-1}(0) + \kappa_0 \|(u, v, w)\|\mathbb{B}, \quad \forall (u, v, w) \in \mathcal{V}. \tag{3.4}$$

Since the function f is strongly convex on any compact subset of $\mathbb{S}^n_{++}$, we know that the KKT point of problem (3.1) is unique. The global Lipschitz continuity of proximal mappings Prox_f and Prox_p implies that the multifunction R is globally continuous. Therefore, there exists a neighborhood $\mathcal{V}_0$ of the origin such that for any $(u, v, w) \in \mathcal{V}_0$ there exists $(x, y, z) \in R^{-1}(u, v, w)$, that is

$$\begin{pmatrix} x - u - \underset{f}{\text{Prox}}\,(x - z - \hat{s}) \\ y - v - \underset{p}{\text{Prox}}(y + z) \\ x - y - w \end{pmatrix} = 0.$$

This, together with the definition of $\mathcal{K}_{pert}$, implies that

$$\mathcal{K}_{pert}((x - u, y - v, z); (u, v, w + u - v)) = 0.$$

Consequently, it holds that $\mathsf{Sol}(u, v, w + u - v) = \{(x - u, y - v, z)\}$. Let $(\bar{x}, \bar{y}, \bar{z})$ be the KKT point of problem (3.1), then $R(\bar{x}, \bar{y}, \bar{z}) = 0$ and $\mathsf{Sol}(0, 0, 0) = \{(\bar{x}, \bar{y}, \bar{z})\}$.

Set $\mathcal{V} = \mathcal{V}_0 \cap \{(u, v, w) : (u, v, w + u - v) \in \mathcal{N}\} \neq \emptyset$ with $\mathcal{N}$ being described in Proposition 2.1. Then, it follows from Proposition 2.1 that

$$\|\mathrm{Sol}(u, v, w + u - v) - \mathrm{Sol}(0, 0, 0)\| = \|(x - u, y - v, z) - (\bar{x}, \bar{y}, \bar{z})\| \leq \kappa \|(u, v, w + u - v)\|.$$

Therefore, for any $(u, v, w) \in \mathcal{V}$, one has

$$\begin{aligned} \|(x, y, z) - (\bar{x}, \bar{y}, \bar{z})\| &\leq \kappa \|(u, v, w + u - v)\| + \|(u, v, 0)\| \\ &\leq (2\kappa + 1)\|(u, v, w)\|. \end{aligned}$$

This implies (3.4) holds. The proof is completed. □

Theorem 3.1 *The infinite sequence* $\{(x^k, y^k, z^k)\}$ *generated by Algorithm 1 converges globally and linearly to a KKT point of problem* (3.1).

Proof Since Algorithm 1 is essentially a special case of the semi-proximal ADMM studied in Han et al. (2018), we can obtain the conclusion directly from Proposition 3.1 and Han et al. (2018, Theorem 2). □

3.2 Implementation details

To implement the algorithm efficiently, we give the details for implementation:

– Closed-form expression of y^{k+1}: from the definition of the augmented Lagrangian function, one has

$$y^{k+1} = \arg\min_y \mathbb{L}_\sigma\left(x^k, y, z^k\right) = \arg\min_y p(y) + \frac{\sigma}{2}\left\|y - x^k - z^k/\sigma\right\|^2.$$

Specifically, we have

(a) if $p(x) = \lambda\|x\|_{1,\mathrm{off}}$, then $y^{k+1} = \mathrm{Prox}_{\sigma^{-1}\lambda\|\cdot\|_{1,\mathrm{off}}}(x^k + z^k/\sigma)$;
(b) if $p(x) = \|x\|_{1,\mathrm{off}} + \mathbb{I}_{\mathcal{C}}(x)$, then by some elementary calculation and Lemma 2.1 , one has

$$y^{k+1} = \sigma^{-1}\mathrm{Prox}_{p_\sigma}\left(\sigma x^k + z^k\right), \quad p_\sigma(x) := \Pi_{\mathcal{C}_\sigma} \circ \mathrm{Prox}_{\lambda\|\cdot\|_{1,\mathrm{off}}}(x),$$

where $\mathcal{C}_\sigma := \{x : \|x\|_2 \leq \sigma\alpha\}$.

– Closed-form expression of x^{k+1}: from the definition of the augmented Lagrangian function, one has

$$x^{k+1} = \arg\min_{x \succ 0} f(x) + \frac{\sigma}{2}\left\|x - y^{k+1} + (z^k + \hat{s})/\sigma\right\|^2 = \mathrm{Prox}_{\sigma^{-1}f}(y^{k+1} - \left(z^k + \hat{s}\right)/\sigma).$$

Specifically, let $\tilde{x} := y^{k+1} - (z^k + \hat{s})$ have the eigenvalue decomposition $\tilde{x} = PDP^T$ with $D = \mathrm{diag}(d)$, from Wang et al. (2010)[Lemma 3.1], one has

$$x^{k+1} = P\mathrm{diag}\Big(\phi_\gamma^+(d)\Big)P^T,\ \ \gamma := \sigma^{-1},$$

where $[\phi_\gamma^+(d)]_i = (\sqrt{d_i^2 + 4\gamma} - d_i)/2,\ i = 1, \ldots, p.$

– Stopping criteria: based on the KKT condition for problem (3.1) and Proposition 3.1, we terminate the algorithm when the relative KKT residual is less than 10^{-5} or the maximum number of iteration 20,000 is attained. Here, the relative KKT residual is given as

$$\eta = \{\eta_p, \eta_d, \eta_c\},$$

where $\eta_p = \frac{\|x-y\|}{1+\|x\|},\ \eta_d = \frac{\|x-\mathrm{Prox}_f(x-z-\hat{s})\|}{1+\|x\|}$, and $\eta_c = \frac{y-\mathrm{Prox}_p(y+z)}{1+\|y\|}$.

4 Numerical experiments

In this section, we evaluate the performance of the proposed model (1.2) by comparing with the sparse Gaussian graphical model (sGGM) with positive semidefinite covariance estimation and the sGGM with IPW estimator. Meanwhile, we use ADMM to solve the three models. Specifically, we consider the following problems:

$$\min_{x\succ 0} f(x) + \langle \hat{s}, x\rangle + p(x), \tag{4.1}$$

where the covariance estimation $\hat{s}$ and the regularizer $p : \mathbb{S}^n \to \mathfrak{R} \cup \{+\infty\}$ are taken respectively as the following forms: given an IPW estimator $s \in \mathbb{S}^n$,

– for sGGM with positive semidefinite covariance estimation (sGGM in short): $p(x) := \lambda\|x\|_{1,\mathrm{off}}$ and $\hat{s} = \Pi_{\mathbb{S}^n_+}(s)$. More specifically, let the given estimator s admit the eigenvalue decomposition $s = U\mathrm{Diag}(\mu_1, \ldots, \mu_n)U^T$ with U being an orthogonal matrix, then $\hat{s} = U\mathrm{Diag}(\hat{\mu}_1 \ldots, \hat{\mu}_n)U^T,\ \hat{\mu}_i = \max\{\mu_i, 0\}, i = 1, \ldots, n$. Here, for a given vector μ, $Diag(\mu)$ is the diagonal matrix whose diagonal elements are the components of μ.
– for sGGM with IPW estimator (rGGM0 in short): $p(x) := \lambda\|x\|_{1,\mathrm{off}}$ and $\hat{s} = s$;
– for rGGM: $p(x) := \lambda\|x\|_{1,\mathrm{off}} + \mathbb{I}_{\mathcal{C}}(x)$ and $\hat{s} = s$.

In this part, the IPW estimator $s \in \mathbb{S}^n$ is obtained from Kolar and Xing (2012). Given a set of partially observed samples $\{\xi_i \in \mathbb{R}^n,\ i = 1, \ldots, m\}$. Let $\delta \in \mathfrak{R}^{m\times n}$ be the matrix with elements given by

$$\delta_{ij} = \begin{cases} 1, & \text{if} \xi_{ij} \text{is observed;} \\ 0, & \text{otherwise.} \end{cases}$$

Then, the (i, j)-th element of the IPW estimator s is taken as

$$s_{ij} = \frac{\sum_{k=1}^{m} \delta_{ki}\delta_{kj}\left(\xi_{ki} - \hat{\mu}_i\right)\left(\xi_{kj} - \hat{\mu}_j\right)}{\sum_{k=1}^{m} \delta_{ki}\delta_{kj}}, \quad i,j = 1, \ldots, n,$$

where $\hat{\mu}_i = (\sum_{k=1}^{m} \delta_{ki})^{-1} \sum_{k=1}^{m} \delta_{ki}\xi_{ki}$.

The performance of the estimations of precision matrix from different methods are evaluated by F1-Score:

$$\mathrm{F1} = \frac{2 \times \mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}},$$

where

$$\mathrm{precision} = \frac{|\hat{x} \cap x^*|}{\text{the number of nonzeros in } \hat{x}}, \quad \mathrm{recall} = \frac{|\hat{x} \cap x^*|}{\text{the number of nonzeros in } x^*}.$$

Here, $|\hat{x} \cap x^*|$ is the number of nonzeros in x^* estimated as nonzeros in $\hat{x}$, $\hat{x}$ and x^* are the estimated precision matrix from the problem (4.1) and the true precision matrix, respectively. The regularized parameter λ and α are chosen by using the modified BIC criterion [see e.g. Städler and Bühlmann (2012)]:

$$(\lambda^*, \alpha^*) := \min_{\lambda,\alpha} m(-\log\det \hat{x}(\lambda,\alpha) + \langle \hat{s}, \hat{x}(\lambda,\alpha)\rangle) + \log(m) \sum_{i \le j} \mathbf{I}_{\{\hat{x}(\lambda,\alpha)_{ij} \neq 0\}}, \tag{4.2}$$

where $\hat{x}(\lambda,\alpha)$ is the optimal solution of problem (4.1), and $\mathbf{I}_{\{\hat{s}_{ij} \neq 0\}} = 1$ if $\hat{s}_{ij} \neq 0$, and 0 otherwise.

4.1 Simulation 1

The data generating processes in this part are from Kolar and Xing (2012), Städler and Bühlmann (2012), Rothman et al. (2008). Assume that $\xi_1, \ldots, \xi_m$, $i.i.d. \sim \mathcal{N}(0, s)$ with

Model 1: $\sigma_{ij} = 0.7^{|i-j|}$, $i,j = 1, \ldots, n$;

Model 2: for $i,j = 1, \ldots, n$, the (i,j)-th element of covariance matrix Σ:

$$\sigma_{ij} = \mathbf{I}_{\{i=j\}} + 0.4\mathbf{I}_{\{|i-j|=1\}} + 0.2\mathbf{I}_{\{|i-j|=2\}} + 0.2\mathbf{I}_{\{|i-j|=3\}} + 0.1\mathbf{I}_{\{|i-j|=4\}},$$

where $\mathbf{I}_{\{i=j\}}$ represents a function which is 1 if $i = j$ and 0 otherwise;

Model 3: $\Omega = B + \gamma I$, where each off-diagonal entry of B is generated independently and equals 0.5 with probability $\zeta = 0.1$ or 0 with probability $1 - \zeta$. Diagonal entries of B are zero, and γ is chosen such that the condition number of $\Omega = \Sigma^{-1}$ is n.

For each model, we obtain the training datasets by deleting completely at random $r\%(r = 10, 20, 30)$ of the synthetic data. It has been pointed out in Rothman et al. (2008) that in *Model 1* and *Model 2*, the number of nonzeros in s^{-1} is proportional to m, whereas in *Model 3*, the number of nonzeros in s^{-1} is proportional to m^2. That is, the precision matrices generated by *Model 1* and *Model 2* have strong sparsity and that generated by *Model 3* have weak sparsity.

We evaluate the performance of rGGM (1.2) by comparing with that of the sGGM and rGGM0. The parameters chosen to obtain the precision matrix estimation are selected by using (4.2). Specifically, the optimal α is selected from set $\{10, 9, 8, \ldots, 1\}$, the optimal regularized parameters λ for *Model 1*, *Model 2*, and *Model 3* are from {0.5, 0.45, ..., 0.35}, {0.2, 0.15, ..., 0.05}, and {0.075, 0.0725, ..., 0.06}, respectively. The Average performance results based on 50 simulation runs are presented in Table 1. Meanwhile, we test the convergence time of ADMM used in rGGM. Specifically, the models with $m = 100, n = 100$ take less than 0.15 s, the models with $m = 150, n = 200$ take no more than 0.4 s, and the models with $m = 200, n = 500$ take less than 4 s. From the table, we can see that the F1-Scores for all three methods tend to decrease with the increase of missing rates for most cases.

For *Model 1*, which has the most strong sparsity of the three data generating processes, the F1-Score obtained by rGGM is significantly better than that generated by the other two methods. The main reason is that the number of nonzeros in the precision matrix obtained from rGGM is in good agreement with that of the true precision matrix. This can be observed from the values of precision and recall corresponding to each estimation method. According as the sparsity decreases, the advantage of the F1-Score obtained by rGGM goes down. However, for the highest dimensional cases with three different data generating processes, the F1-Score obtained by rGGM is still better than that generated by the other two methods. For better illustration, we also give the boxplots (Fig. 1) of the data generated by three models with $m = 200, n = 500, r = 30$.

4.2 Simulation 2: protein

In this part, we generate the data based on the real dataset *protein*[1] ($m = 6621$, $n = 357$) which can be obtained from LIBSVM library (Chang and Lin 2011). The dataset *protein* contains three classes and the sample sizes corresponding to each class are 3112, 2323, and 1186, respectively.

The datasets used in this part are generated by the following steps. For each class, we first generate the sample covariance matrix $s_i \in \mathbb{S}_+^{357} (i = 1, 2, 3)$ based on the complete data from dataset *protein* and select edges by using the sparse Gaussian graphical model (1.1). The selected edges will be taken as the "true" edges. Then, we randomly generate 50 datasets with samples from multivariate Gaussian distribution $\mathcal{N}(0, s_i)$. For each synthetic dataset, we further produce completely at random $r\%$ missing values, $r = 10, 20, 30$. Finally, we evaluate the performances of sGGM, rGGM0, and rGGM with F1-Score based on the 50 synthetic datasets.

The numerical results are visualized by box plots, see Fig. 2. We can see from the figure that the F1-Score obtained by rGGM is higher than that of the other two methods for 6 out of 9 cases. For the other 3 cases, the F1-Score of rGGM is still comparable with that of the other two methods.

[1] Available at: https://www.csie.ntu.edu.tw/~cjlin/libsvm .

Table 1 Comparison of average precision, recall, and F1-Score for sGGM, rGGM0, and rGGM with data generated by Model 1, Model 2, and Model 3 based on 50 simulation runs

Model		Precision			Recall			F1		
(m, n)	r	sGGM	rGGM0	rGGM	sGGM	rGGM0	rGGM	sGGM	rGGM0	rGGM
Model1 (100,100)	10	0.650	0.654	0.726	0.995	0.996	0.979	0.786	0.789	0.834
	20	0.594	0.590	0.716	0.990	0.992	0.961	0.742	0.740	0.821
	30	0.526	0.504	0.715	0.973	0.977	0.907	0.683	0.665	0.799
Model1 (150,200)	10	0.639	0.647	0.724	1.000	1.000	0.990	0.780	0.786	0.836
	20	0.617	0.625	0.745	0.999	0.999	0.974	0.763	0.769	0.844
	30	0.575	0.565	0.753	0.995	0.997	0.948	0.729	0.721	0.839
Model1 (200,500)	10	0.616	0.624	0.794	1.000	1.000	0.989	0.763	0.769	0.881
	20	0.607	0.618	0.784	1.000	1.000	0.979	0.755	0.764	0.871
	30	0.584	0.581	0.767	0.999	0.999	0.963	0.737	0.735	0.854
Model2 (100,100)	10	0.533	0.525	0.522	0.195	0.197	0.206	0.286	0.287	0.296
	20	0.474	0.456	0.453	0.213	0.218	0.230	0.294	0.295	0.305
	30	0.423	0.285	0.399	0.228	0.606	0.257	0.296	0.388	0.312
Model2 (150,200)	10	0.538	0.518	0.525	0.171	0.172	0.186	0.259	0.258	0.275
	20	0.427	0.393	0.396	0.182	0.187	0.202	0.255	0.254	0.267
	30	0.346	0.157	0.309	0.194	0.544	0.223	0.249	0.244	0.259
Model2 (200,500)	10	0.510	0.475	0.493	0.161	0.162	0.180	0.244	0.241	0.264
	20	0.365	0.315	0.326	0.167	0.171	0.190	0.230	0.222	0.240
	30	0.249	0.072	0.210	0.178	0.454	0.207	0.208	0.124	0.208
Model3 (100,100)	10	0.349	0.350	0.255	0.696	0.700	0.768	0.465	0.467	0.383
	20	0.301	0.296	0.232	0.659	0.659	0.749	0.414	0.408	0.355
	30	0.255	0.210	0.209	0.620	0.650	0.730	0.361	0.317	0.325

Table 1 (continued)

Model		Precision			Recall			F1		
(m, n)	r	sGGM	rGGM0	rGGM	sGGM	rGGM0	rGGM	sGGM	rGGM0	rGGM
Model3 (150,200)	10	0.406	0.410	0.258	0.474	0.479	0.565	0.437	0.442	0.354
	20	0.356	0.358	0.242	0.449	0.457	0.557	0.397	0.401	0.338
	30	0.305	0.292	0.222	0.430	0.445	0.552	0.357	0.352	0.317
Model3 (200,500)	10	0.376	0.384	0.236	0.202	0.204	0.294	0.263	0.266	0.262
	20	0.341	0.352	0.225	0.200	0.200	0.299	0.252	0.255	0.257
	30	0.301	0.308	0.208	0.199	0.201	0.311	0.239	0.243	0.249

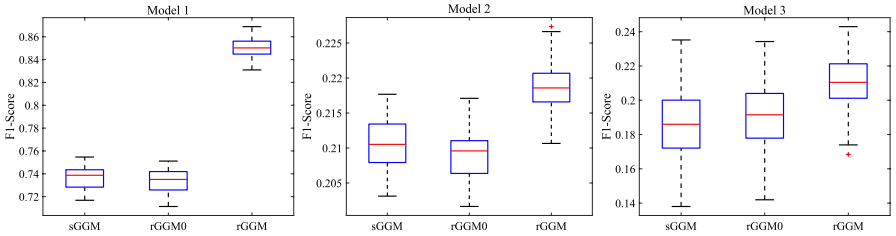

Fig. 1 Comparison of F1-Score for sGGM, rGGM0, and rGGM with data generated by *Model 1*, *Model 2*, and *Model 3* based on 50 simulation runs ($m = 200, n = 500, r = 30$)

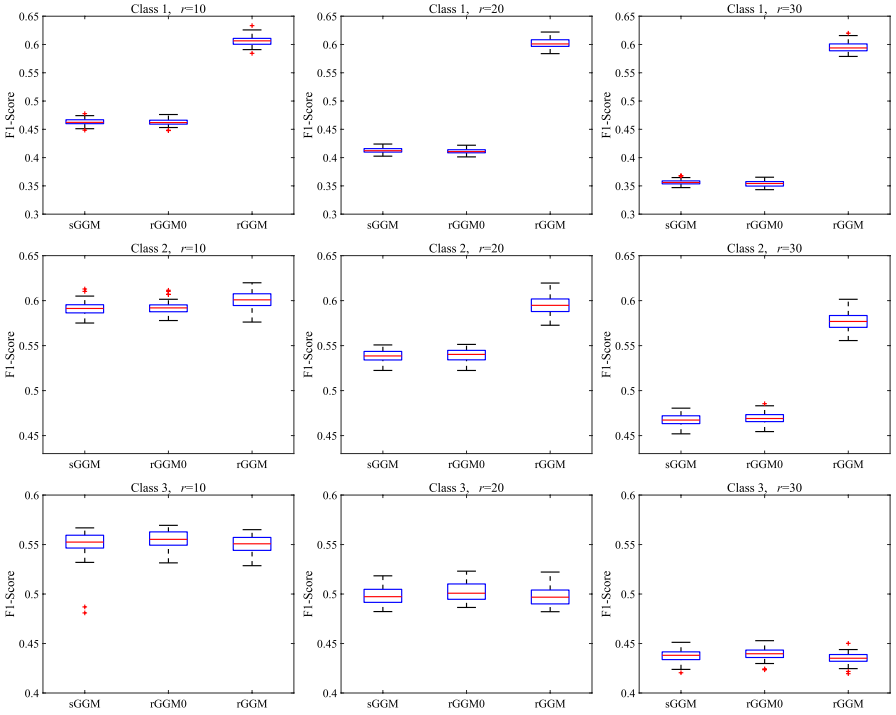

Fig. 2 Comparison of F1-Score for sGGM, rGGM0, and rGGM with 3 class datasets generated from *protein*. Here, the sample sizes corresponding to each class are 3112, 2323, and 1186, respectively

4.3 Simulation 3: university webpages

For illustration, we also explore the performance of rGGM on the dataset *University Webpages*, which was originally collected by the World Wide Knowledge Base (Web–>Kb) project of the CMU text learning group.[2] In this part, we use the

[2] Available at: http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/.

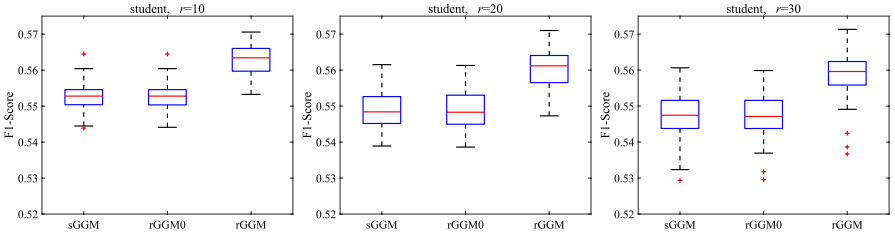

Fig. 3 Comparison of F1-Score for sGGM, rGGM0, and rGGM with class *student* generated from the data set *University Webpages*

pre-processed dataset *webkb-test-stemmed*,[3] which contains four different classes: student, staff, course, and project. More specifically, we select the class *student*, the one has the largest sample size $m = 544$. The pre-precessing details can be found in Cardoso-Cachopo (2007, Section 2.8). We further process the dataset and obtain the term-document matrix $X \in \mathfrak{R}^{544\times n}$ by log-entropy weighting method [see e.g. Dumais (1991), Guo et al. (2011)]. Here, we take $n = 300$.

Since the term-document matrix $X \in \mathfrak{R}^{544\times 300}$ is very sparse (the percentage of nonzeros is 10.1%), we produce completely at random r% nonzero missing values ($r = 10, 20, 30$) and evaluate the performances of sGGM, rGGM0, and rGGM with F1-Score based on the 50 synthetic datasets. The results are visualized in Fig. 3. As shown in the figure, the F1-scores for all the methods tend to decrease as the missing rate increases. But, the rGGM still has better F1-Score on the class *student* of the dataset *University Webpages*.

5 Conclusion

In this paper, we proposed an estimator of sparse precision matrix based on the dataset with partially observed observations. The estimator was obtained by using a robust convex optimization problem that can be solved by a linearly convergent first-order method. The numerical results showed that the presented estimators usually have satisfactory F1-Scores. It is commonly accepted that the nonconvex regularizers have better performance than convex regularizers. Therefore, we will design the efficient numerical algorithms for solving the robust Gaussian graphical model with nonconvex regularizers based on the theoretical analysis of the convex counterpart in the future.

[3] Available at: https://ana.cachopo.org/datasets-for-single-label-text-categorization.

Appendix

Proof of Lemma 2.1

Lemma 6.1 (*Yu* 2013, *Theorem* 1) *Let f and g be two closed convex proper functions. A sufficient condition for* $\mathrm{Prox}_{f+g} = \mathrm{Prox}_f \circ \mathrm{Prox}_g$ *is*

$$\partial g(x) \subseteq \partial g\big(\mathrm{Prox}_f(x)\big), \ \forall x$$

Let $p : \mathbb{S}^n \to \mathbb{R} \cup \{+\infty\}$ be defined by

$$p(x) = \lambda \|x\|_{1,\mathrm{off}} + \varphi(x), \ \varphi(x) := \mathbb{I}_{\mathcal{C}}(x), \ \mathcal{C} := \{x : \ \|x\| \le \alpha\}. \tag{6.1}$$

It follows from the definition of proximal mapping that

$$\mathrm{Prox}_{\varphi}(x) = \Pi_{\mathcal{C}}(x) = \begin{cases} x, & \|x\| \le \alpha, \\ \dfrac{\alpha x}{\|x\|}, & \|x\| > \alpha. \end{cases} \tag{6.2}$$

Lemma 6.2 *Let function p be defined by* (6.1), *it holds that*

$$\mathrm{Prox}_p = \Pi_{\mathcal{C}} \circ \mathrm{Prox}_{\lambda\|\cdot\|_{1,\mathrm{off}}}.$$

Proof From Lemma 6.1, it is sufficient to show that

$$\partial \mathrm{Prox}_{\lambda\|\cdot\|_{1,\mathrm{off}}}(x) \subseteq \partial \mathrm{Prox}_{\lambda\|\cdot\|_{1,\mathrm{off}}}(y), \ y = \Pi_{\mathcal{C}}(x), \ \forall x \in \mathbb{R}^m. \tag{6.3}$$

From (6.2), it is sufficient to consider the following cases:

(a) If $\|x\| \le \alpha$, then $y = x$. Therefore, the relationship (6.3) holds.
(b) If $\|x\| > \alpha$, then $y = \alpha x/\|x\|$, which means $\mathrm{sgn}(y) = \mathrm{sgn}(x)$. Therefore, the relationship (6.3) also holds in this case.

The proof is completed. □

Proof of Proposition 2.1

Lemma 6.3 (*Zhang et al.* 2020, *Lemma* 3.1) *Let $f(x) := -\log\det x$. Then all $\mathcal{G}_f \in \partial \mathrm{Prox}_f(Z)$ are self-adjoint and positive definite with $\lambda_{\max}(\mathcal{G}_f) < 1$.*

Lemma 6.4 *Let $x \in \mathbb{S}^n$ and $\mathcal{B} : \mathbb{S}^n \to \mathbb{S}^n$ be any self-adjoint positive definite operator, p is the function defined in Lemma2.1. Then, for any chosen $\mathcal{G}_p \in \partial \mathrm{Prox}_p(x)$, the linear operator $I - \mathcal{G}_p + \mathcal{G}_p\mathcal{B}$ is nonsingular.*

Proof It follows Lemma 6.1 that Prox_p is the projection onto the closed convex set $\mathcal{C}$. Therefore, we know from Sun and Qi (2001, Theorem 2.3) that any element

$\mathcal{G}_p \in \partial \mathrm{Prox}_p(x)$ is self-adjoint, positive definite, and $\lambda_{\max}(\mathcal{G}_p) \in [0, 1]$. The proof can be completed by Zhang et al. (2020, Lemma 3.2). □

Lemma 6.5 *Let $\mathcal{K}_{pert}$ be the KKT mapping defined by (2.3), $(\bar{x}, \bar{y}, \bar{z})$ be the KKT point of problem (1.4). Then, Any element in $\partial_{(x,y,z)}\mathcal{K}_{pert}((\bar{x}, \bar{y}, \bar{z}), (0, 0, 0))$ is nonsingular.*

Proof Since Prox_p is directionally differentiable, it follows from the chain rule presented in Sun (2006, Lemma 2.1) that for any $\mathcal{G} \in \partial_{(x,y,z)}\mathcal{K}_{pert}((\bar{x}, \bar{y}, \bar{z}), (0, 0, 0))$, there exist $\mathcal{G}_f \in \partial \mathrm{Prox}_f(\bar{x} - \bar{z} - \hat{s})$ and $\mathcal{G}_p \in \partial \mathrm{Prox}_p(\bar{y} + \bar{z})$ such that

$$\mathcal{G}(\Delta x, \Delta y, \Delta z) = \begin{pmatrix} \Delta x - \mathcal{G}_f(\Delta z + \Delta x) \\ \Delta y - \mathcal{G}_p(\Delta y + \Delta z) \\ \Delta x - \Delta y \end{pmatrix}, \quad \forall\, (\Delta x, \Delta y, \Delta z) \in \mathbb{S}^n \times \mathbb{S}^n \times \mathbb{S}^n.$$

Suppose that there exists $(\Delta x, \Delta y, \Delta z) \in \mathbb{S}^n \times \mathbb{S}^n \times \mathbb{S}^n$ such that $\mathcal{G}(\Delta x, \Delta y, \Delta z) = 0$, i.e.,

$$\begin{cases} \Delta x - \mathcal{G}_f(\Delta z + \Delta x) = 0, \\ \Delta y - \mathcal{G}_p(\Delta y + \Delta z) = 0, \\ \Delta x - \Delta y = 0. \end{cases} \tag{6.4}$$

It follows from Lemma 6.3 that both $\mathcal{G}_f$ and $\mathcal{G}_f^{-1} - I$ are self-adjoint and positive definite. This, together with (6.4), implies that

$$\Delta z = \left(\mathcal{G}_f^{-1} - I\right)\Delta x \ \text{ and } \ \left(I - \mathcal{G}_p + \mathcal{G}_p\left(\mathcal{G}_f^{-1} - I\right)\right)\Delta x = 0. \tag{6.5}$$

We know from Lemma 6.4 that $(I - \mathcal{G}_p + \mathcal{G}_p(\mathcal{G}_f^{-1} - I))$ is nonsingular. This, together with (6.5), implies that

$$\Delta x = 0, \ \ \Delta y = 0, \ \text{ and } \ \Delta z = 0.$$

Therefore, $\mathcal{G}$ is nonsingular. The proof is completed. □

In order to give the proof of Proposition 2.1, we recall the implicit theorem from Clarke et al. (1998). Let $\mathbb{X}$ be a Hilbert space and $\mathbb{M}$ be a metric space. Consider the equation

$$\mathcal{H}(x, \alpha) = 0,$$

where $\mathcal{H}$ is a mapping from $\mathbb{X} \times \mathbb{M}$ to $\mathbb{X}$. Assume that $V \subseteq \mathbb{X}$ is an open set such that $\mathcal{H}$ is continuous on $V \times \mathbb{M}$ and such that the partial derivative $\partial_x \mathcal{H}(x, \alpha)$ exists for all $(x, \alpha) \in V \times \mathbb{M}$, and is continuous jointly in $(x, \alpha) \in V \times \mathbb{M}$.

The following result is from Clarke et al. (1998, Theorem 3.6), which is usually named as *Clarke's implicit function theorem*.

Lemma 6.6 *Let $(x_0, \alpha_0) \in V \times \mathbb{M}$ be a point satisfying $\mathcal{H}(x_0, \alpha_0) = 0$. Then one has*

(a) *If* $\partial_x \mathcal{H}(x_0, \alpha_0)$ *is onto and one to one, then there exist neighborhoods* $\mathcal{N}_x$ *of* x_0 *and* $\mathcal{N}_\alpha$ *of* α_0 *and a unique continuous function* $\hat{x}(\cdot) : \mathcal{N}_\alpha \to \mathcal{N}_x$ *with* $\hat{x}(\alpha_0) = x_0$ *such that* $\mathcal{H}(\hat{x}(\alpha), \alpha) = 0,\ \forall \alpha \in \mathcal{N}_\alpha$.

(b) *If in addition* $\mathcal{H}$ *is Lipschitz in a neighborhood of* (x_0, α_0), *then* $\hat{x}$ *is Lipschitz.*

Now, we are ready to present the proof of Proposition 2.1.

Proof The global Lipschitz continuities of the proximal mappings Prox_f and Prox_p imply that the mapping $\mathcal{K}_{pert}$ defined by (2.3) is Lipschitz continuous. Therefore, the proof can be completed by Lemmas 6.5, 6.6, and the fact that for any (u, v, w), the set $\text{Sol}(u, v, w)$ must be a singleton if it is nonempty. □

References

Ahn M, Pang J-S, Xin J (2017) Difference-of-convex learning: directional stationarity, optimality, and sparsity. SIAM J Optim 27(3):1637–1665

Cardoso-Cachopo A (2007) Improving methods for single-label text categorization. PhD Thesis, Instituto Superior Tecnico, Universidade Tecnica de Lisboa

Chang C-C, Lin C-J (2011) Libsvm: a library for support vector machines. ACM Trans Intell Syst Technol (TIST) 2(3):27

Clarke FH, Ledyaev YS, Stern RJ, Wolenski PR (1998) Nonsmooth analysis and control theory, vol 178. Springer, New York

Dempster AP (1972) Covariance selection. Biometrics 28(1):157–175

Dontchev A, Rockafellar R (2004) Regularity and conditioning of solution mappings in variational analysis. Set-Valued Anal 12(1–2):79–109

Dumais ST (1991) Improving the retrieval of information from external sources. Behav Res Methods Instrum Comput 23(2):229–236

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96(456):1348–1360

Fan J, Liao Y, Liu H (2016) An overview of the estimation of large covariance and precision matrices. Economet J 19(1):C1–C32

Fan R, Jang B, Sun Y, Zhou S (2019) Precision matrix estimation with noisy and missing data. In: The 22nd international conference on artificial intelligence and statistics, vol 89, pp 2810–2819. PMLR

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

Guo J, Levina E, Michailidis G, Zhu J (2011) Joint estimation of multiple graphical models. Biometrika 98(1):1–15

Han D, Sun D, Zhang L (2018) Linear rate convergence of the alternating direction method of multipliers for convex composite programming. Math Oper Res 43(2):622–637

Hsieh CJ, Sustik MA, Dhillon IS, Ravikumar P (2014) QUIC: quadratic approximation for sparse inverse covariance estimation. J Mach Learn Res 15:2911–2947

Kolar M, Xing EP (2012) Estimating sparse precision matrices from data with missing values. In: Proceedings of the 29th international conference on machine learning, Edinburgh, Scotland, UK, pp 635–642

Loh P-L, Wainwright MJ (2015) Regularized m-estimators with nonconvexity: statistical and algorithmic theory for local optima. J Mach Learn Res 16(1):559–616

Lounici K (2014) High-dimensional covariance matrix estimation with missing observations. Bernoulli 20(3):1029–1058

Lu L, Toh KC (2010) An inexact interior point method for L1-regularized sparse covariance selection. Math Program Comput 2(3–4):291–315

Moreau J-J (1965) Proximité et dualité dans un espace Hilbertien. Bull Soc Math France 93(2):273–299

Park S, Lim J (2019) Non-asymptotic rate for high-dimensional covariance estimation with non-independent missing observations. Stat Probab Lett 153:113–123

Park S, Wang X, Lim J (2020) Estimating high-dimensional covariance and precision matrices under general missing dependence. arXiv preprint arXiv:2006.04632
Pavez E, Ortega A (2021) Covariance matrix estimation with non uniform and data dependent missing observations. IEEE Trans Inf Theory 67(2):1201–1215
Rockafellar TR, Wets RJ-B (1998) Variational analysis. Sobolev BV Sp MPS-SIAM Ser Optim 30:324–326
Rothman AJ, Bickel PJ, Levina E, Zhu J (2008) Sparse permutation invariant covariance estimation. Electron J Stat 2:494–515
Seaman SR, White IR (2013) Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res 22(3):278–295
Städler N, Bühlmann P (2012) Missing values: sparse inverse covariance estimation and an extension to sparse regression. Stat Comput 22(1):219–235
Sun DF (2006) The strong second-order sufficient condition and constraint nondegeneracy in nonlinear semidefinite programming and their implications. Math Oper Res 31(4):761–776
Sun DF, Qi L (2001) Solving variational inequality problems via smoothing-nonsmooth reformulations. J Comput Appl Math 129(1–2):37–62
Sun S, Huang R, Gao Y (2012) Network-scale traffic modeling and forecasting with graphical lasso and neural networks. J Transp Eng 138(11):1358–1367
Tang P, Wang C, Sun D, Toh K-C (2020) A sparse semismooth Newton based proximal majorization-minimization algorithm for nonconvex square-root-loss regression problems. J Mach Learn Res 21(226):1–38
Wang C, Sun D, Toh K-C (2010) Solving log-determinant optimization problems by a Newton-CG primal proximal point algorithm. SIAM J Optim 20(6):2994–3013
Wang T, Ren Z, Ding Y, Fang Z, Sun Z, MacDonald ML, Sweet RA, Wang J, Chen W (2016) Fastggm: an efficient algorithm for the inference of gaussian graphical model in biological networks. PLoS Comput Biol 12(2):e1004755
Ye JJ, Ye XY (1997) Necessary optimality conditions for optimization problems with variational inequality constraints. Math Oper Res 22(4):977–997
Yosida K (1964) Functional analysis. Springer, Berlin
Yu Y-L (2013) On decomposing the proximal map. In: Proceedings of advances in neural information processing systems, pp 91–99
Yuan X (2012) Alternating direction method for covariance selection models. J Sci Comput 51(2):261–273
Yuan M, Lin Y (2007) Model selection and estimation in the Gaussian graphical model. Biometrika 94(1):19–35
Yuan X, Zeng S, Zhang J (2020) Discerning the linear convergence of admm for structured convex optimization through the lens of variational analysis. J Mach Learn Res 21:1–75
Zerenner T, Friederichs P, Lehnertz K, Hense A (2014) A gaussian graphical model approach to climate networks. Chaos 24(2):023103
Zhang C-H (2010) Nearly unbiased variable selection under minimax concave penalty. Ann Stat 38(2):894–942
Zhang A, Fang J, Liang F, Calhoun VD, Wang Y-P (2018) Aberrant brain connectivity in schizophrenia detected via a fast gaussian graphical model. IEEE J Biomed Health Inf 23(4):1479–1489
Zhang Y, Zhang N, Sun D, Toh KC (2020) A proximal point dual newton algorithm for solving group graphical lasso problems. SIAM J Optim 30(3):2197–2220