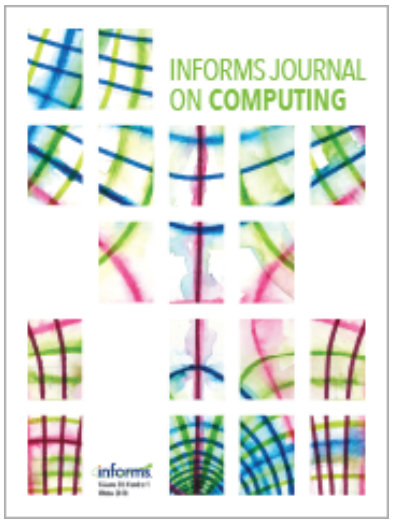


INFORMS Journal on Computing

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

On the Solution of ℓ_0-Constrained Sparse Inverse Covariance Estimation Problems

Dzung T. Phan, Matt Menickelly

To cite this article:
Dzung T. Phan, Matt Menickelly (2020) On the Solution of ℓ_0-Constrained Sparse Inverse Covariance Estimation Problems. INFORMS Journal on Computing

Published online in Articles in Advance 08 Oct 2020

. https://doi.org/10.1287/ijoc.2020.0991

http://pubsonline.informs.org/journal/ijoc

INFORMS JOURNAL ON COMPUTING
Articles in Advance, pp. 1–20
ISSN 1091-9856 (print), ISSN 1526-5528 (online)

On the Solution of ℓ_0-Constrained Sparse Inverse Covariance Estimation Problems

Dzung T. Phan,[a] Matt Menickelly[b]

[a] IBM T.J. Watson Research Center, Yorktown Heights, New York 10598; [b] Argonne National Laboratory, Lemont, Illinois 60439
Contact: phandu@us.ibm.com, https://orcid.org/0000-0003-1579-7035 (DTP); mmenickelly@anl.gov (MM)

Received: March 11, 2019
Revised: December 19, 2019; April 15, 2020; April 28, 2020
Accepted: May 1, 2020
Published Online in Articles in Advance: October 8, 2020

https://doi.org/10.1287/ijoc.2020.0991

Abstract. The sparse inverse covariance matrix is used to model conditional dependencies between variables in a graphical model to fit a multivariate Gaussian distribution. Estimating the matrix from data are well known to be computationally expensive for large-scale problems. Sparsity is employed to handle noise in the data and to promote interpretability of a learning model. Although the use of a convex ℓ_1 regularizer to encourage sparsity is common practice, the combinatorial ℓ_0 penalty often has more favorable statistical properties. In this paper, we directly constrain sparsity by specifying a maximally allowable number of nonzeros, in other words, by imposing an ℓ_0 constraint. We introduce an efficient approximate Newton algorithm using warm starts for solving the nonconvex ℓ_0-constrained inverse covariance learning problem. Numerical experiments on standard data sets show that the performance of the proposed algorithm is competitive with state-of-the-art methods.

Summary of Contribution: The inverse covariance estimation problem underpins many domains, including statistics, operations research, and machine learning. We propose a scalable optimization algorithm for solving the nonconvex ℓ_0-constrained problem.

History: Accepted by Antonio Frangioni, Area Editor for Design & Analysis of Algorithms—Continuous.
Supplemental Material: Data are available at https://doi.org/10.1287/ijoc.2020.0991.

Keywords: ℓ_0-constrained • sparsity • gradient projection • approximate Newton • inverse covariance

1. Introduction

Given m independent and identically distributed (i.i.d) data samples $\{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ from an n-variate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we wish to estimate the mean $\boldsymbol{\mu} \in \mathbb{R}^n$ and the covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$. It is well known that $\frac{1}{m}\sum_{i=1}^m \mathbf{y}_i$ is the maximum likelihood estimator (MLE) for $\boldsymbol{\mu}$, whereas the solution of

$$\min_{\mathbf{X}\succ\mathbf{0}} \ \operatorname{tr}(\mathbf{S}\mathbf{X}) - \log\det\mathbf{X} \tag{1}$$

is the MLE for the inverse covariance $\boldsymbol{\Sigma}^{-1}$, where $\mathbf{S} \succeq \mathbf{0}$ is the sample covariance $\mathbf{S} = \frac{1}{m}\sum_{i=1}^m (\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})^T$. The sparsity pattern of $\mathbf{X} = \boldsymbol{\Sigma}^{-1}$ reveals conditional independences between variables; in many practical applications, the inverse covariance is sparse. In general, however, $\mathbf{X}$ obtained from (1) is a dense matrix, especially when there is noise in data. Common practice is to explicitly enforce sparsity in $\mathbf{X}$ via some mechanism in the estimation model.

A *convex* model based on the ℓ_1-regularized maximum likelihood problem

$$\min_{\mathbf{X}\succ\mathbf{0}} \ \operatorname{tr}(\mathbf{S}\mathbf{X}) - \log\det(\mathbf{X}) + \lambda\|\mathbf{X}\|_1 \tag{2}$$

has been extensively studied; here, $\|\mathbf{X}\|_1 = \sum_{i,j=1}^n |X_{ij}|$ is the component-wise ℓ_1 norm of the matrix $\mathbf{X}$ and $\lambda > 0$ is a regularization parameter to control sparsity. A number of works have been dedicated to the fast solution of (2); for a nonexhaustive, but diverse, array of optimization methods, see d'Aspremont et al. (2008), Friedman et al. (2008), Li and Toh (2010), Scheinberg et al. (2010), Olsen et al. (2012), Hsieh et al (2014), Treister et al. (2016), and Zhang et al. (2018).

In spite of its computationally efficient advantages, the ℓ_1 norm model has several statistical drawbacks. Research has shown that ℓ_1-regularized models can be asymptotically biased (Zou 2006) and can provide arbitrarily worse predictive accuracy than an ℓ_0-model (Johnson et al. 2015), the subject of this paper. The ℓ_0 norm of $\mathbf{X}$, denoted by $\|\mathbf{X}\|_0$, counts the number of nonzeros of the argument. On the other hand, an ℓ_0-based model is often able to recover the sparsity structure better than its ℓ_1-counterpart can (Lu and Zhang 2013, Marjanovic and Hero 2015). Some works (Marjanovic and Hero 2015, Liu et al. 2016, Marjanovic et al. 2016) address minimizing the negative likelihood function with a regularized ℓ_0 penalty. However, imposing sparsity via regularization in the objective function instead of in the constraint set can introduce bias (Bertsimas et al. 2016).

In this paper, we investigate a *nonconvex* ℓ_0-constrained model. In particular, rather than employing an ℓ_1-regularizer to control sparsity as in (2), we define a sparsity constraint set

$$\Omega \triangleq \{\mathbf{X} \in \mathbb{R}^{n\times n} : \|\mathbf{X}\|_0 \le \kappa, \;\; X_{i,j} = 0 \;\; \forall (i,j) \in \Im\},$$

where κ is a maximally allowable number of nonzeros and $\Im$ is the index set of variables known to be conditionally independent. Given Ω, we then directly constrain sparsity by solving

$$\begin{array}{ll} \min\limits_{\mathbf{X}} & \mathrm{tr}(\mathbf{SX}) - \log\det(\mathbf{X}) \\ \text{s.t.} & \mathbf{X} \in \Omega \cap \mathbb{S}^n_{++}, \end{array} \tag{3}$$

where $\mathbb{S}^n_{++}$ is the set of symmetric positive definite matrices in $\mathbb{R}^{n\times n}$. Problem (3) has a nonconvex constraint set; hence, local minimizers are not necessarily global minimizers. Some previous work has attempted to solve (3), particularly via a penalty decomposition (PD) method (Lu and Zhang 2013) and a gradient hard thresholding pursuit algorithm (Yuan et al. 2014, 2018). Although these methods guarantee only the identification of points satisfying necessary optimality conditions, the solutions returned by these methods are often practically satisfying in many applications.

As a main contribution, we develop a two-step approximate Newton algorithm (HANA) with warm starts to solve (3). We extend this method with a homotopy approach, solving a sequence of regularized subproblems controlled by two parameters. For each subproblem, a gradient projection step is used to move closer to a local solution as well as to identify a lower-dimensional working subspace. In this step, feasibility with respect to the sparsity constraints is guaranteed via projection, whereas symmetric positive definiteness is ensured through a linesearch procedure. To improve the empirical convergence rate, we occasionally compute approximate Newton steps in a subspace determined from information obtained during the gradient projection step. A homotopy method applied to ℓ_2-regularized subproblems is used as a wrapper around this proposed method to mitigate the difficulties of the unboundedness of the solution set and the practical concerns of becoming trapped at local minima. We prove convergence of the proposed algorithm to a stationary point of the ℓ_0-constrained Problem (3). We do not prove nonasymptotic convergence rates of the method discussed in this paper, but we will demonstrate that these extensions are indeed beneficial.

The paper is organized as follows. In Section 2, we give preliminary results for the Problem (3). In Section 3, we characterize optimality conditions for ℓ_0-constrained problems and describe our algorithm. In Section 4, we provide convergence analysis for our approximate Newton algorithm. Section 5 numerically compares the performance of the proposed algorithm with other methods. We provide concluding remarks in Section 6.

1.1. Notation

Let $f : \mathbb{R}^{n\times n} \to \mathbb{R}$ be a function of a matrix. For any $\mathbf{X} \in \mathbb{R}^{n\times n}$, $\nabla f(\mathbf{X})$ denotes the gradient of f, which is a matrix in $\mathbb{R}^{n\times n}$. The Hessian of f is denoted by $\nabla^2 f(\mathbf{X})$. We let $\mathrm{tr}(\mathbf{X})$ denote the trace operator, in other words, $\sum_{i=1}^n X_{ii}$. The vectorization operator $\mathrm{vec}(\mathbf{X})$ transforms $\mathbf{X}$ into a vector in $\mathbb{R}^{n^2}$ by stacking the columns from left to right. We denote $\sigma_{\min}(\mathbf{X})$ and $\sigma_{\max}(\mathbf{X})$ as the minimum and maximum eigenvalues of $\mathbf{X}$, respectively. For an index set $\mathcal{F} \subseteq \{1,\dots,n\}\times\{1,\dots,n\}$, $[\mathbf{X}]_{\mathcal{F}}$ represents a matrix obtained from the zero matrix $\mathbf{0}_{n\times n}$ when replacing the (i,j)-th entry by $X_{i,j}$ for every $(i,j) \in \mathcal{F}$. Let $\otimes$ denote the Kronecker product of matrices and let $\mathbf{e}_i$ denote a unit vector with a one in the ith coordinate. We refer to $\mathbf{I}_n$ as the n by n identity matrix. For any finite set $\mathcal{S}$, $|\mathcal{S}|$ denotes the number of elements of $\mathcal{S}$.

2. Preliminary Results

First, we characterize the feasibility condition for the Problem (3).

Lemma 1. *Suppose that there is no $(i,i) \in \Im$ for any $i \in \{1,\dots,n\}$. Then the feasible set $\Omega \cap \mathbb{S}^n_{++}$ of the Problem (3) is nonempty if and only if $\kappa \ge n$.*

Proof. If $\mathbf{X}$ is a feasible solution of (3), then $\mathbf{X} > \mathbf{0}$. We have $X_{ii} = \mathbf{e}_i^{\mathrm{T}}\mathbf{X}\mathbf{e}_i > 0$ where $\mathbf{e}_i$ is a vector of all zeros except for a one in the ith entry. Hence, $\kappa \ge \|\mathbf{X}\|_0 \ge n$. Conversely, if $\kappa \ge n$, then $\mathbf{X} = \mathbf{I}_n$ is a feasible solution of (3). □

From Lemma 1, throughout the paper, we always assume that Assumption 1 holds.

Assumption 1. *The parameter κ is an integer satisfying $\kappa \in [n, n^2)$ and $(i,i) \notin \Im$ for every $i \in \{1,\dots,n\}$.*

Observe that if $\kappa = n^2$, then the problem reduces to an unconstrained convex one. We now give a hardness result, demonstrating that for any $\kappa \ge n$ in (3), there exists no $\lambda > 0$ such that an optimal solution to the ℓ_1-regularized model (2) with regularization parameter λ recovers the global optimum of (3). This suggests an inherent need for solving (3) without resorting to convex formulations.

Theorem 1. *There exists no value of $\lambda > 0$ such that an optimal solution of (2) globally minimizes (3).*

Proof. Toward a contradiction, suppose there exists a pair $\kappa \geq n$ and $\lambda > 0$ such that $\mathbf{X}$ is a global minimizer of both (2) and (3). Define $\mathbf{Y} = \mathbf{X}^{-1}$ and $\psi(\mathbf{X}) = -\log\det(\mathbf{X}) + \operatorname{tr}(\mathbf{SX})$. Because $\mathbf{S}$ is positive, semidefinite, and nonzero, there exists at least one diagonal entry $S_{ii} > 0$. Without loss of generality, suppose $S_{11} > 0$. For any α such that $|\alpha| < \min(X_{11}, \sigma_{\min}(\mathbf{X}))$, we have both $\|\mathbf{X} + \alpha\mathbf{e}_1\mathbf{e}_1^T\|_0 = \|\mathbf{X}\|_0 \leq \kappa$ and $\mathbf{X} + \alpha\mathbf{e}_1\mathbf{e}_1^T$ is positive definite. Additionally, because $(\mathbf{X} + \alpha\mathbf{e}_1\mathbf{e}_1^T)_{kl} = 0$ for any $(k,l) \in \Im$, we conclude that $\mathbf{X} + \alpha\mathbf{e}_1\mathbf{e}_1^T$ is feasible in (3) for every α such that $|\alpha| < \min(X_{11}, \sigma_{\min}(\mathbf{X}))$.

Moreover, using a property of determinants, we have

$$\begin{aligned}\psi\left(\mathbf{X} + \alpha\mathbf{e}_1\mathbf{e}_1^T\right) - \psi(\mathbf{X}) &= -\log(\det(\mathbf{X})(1 + \alpha Y_{11})) \\ &\quad + \operatorname{tr}\left(\mathbf{S}\left(\mathbf{X} + \alpha\mathbf{e}_1\mathbf{e}_1^T\right)\right) \\ &\quad + \log\det(\mathbf{X}) - \operatorname{tr}(\mathbf{SX}) \\ &= -\log(1 + \alpha Y_{11}) + \alpha S_{11} \triangleq g(\alpha).\end{aligned}$$

Note that if $\mathbf{X}$ is a minimizer of (3), then $g(\alpha) \geq 0$ for every $|\alpha| < \min(X_{11}, \sigma_{\min}(\mathbf{X}))$. Because g is a (nonnegative and smooth) convex one-dimensional function of α in the neighborhood of zero and $g(0) = 0$, we conclude that $g'(0) = 0$. This implies that $S_{11} = Y_{11}$. From a characterization of optimal solutions of (2) (see, e.g., the proof of lemma 7 in Hsieh et al. 2014) we have that if $\mathbf{X}$ is optimal for (2), then $S_{11} = Y_{11} - \lambda$. Because $\lambda > 0$, this is a contradiction. □

Definition 1. Consider the optimization problem

$$\begin{aligned}\min_{\mathbf{X}} \quad & f(\mathbf{X}) \\ \text{s.t.} \quad & \mathbf{X} \in \Omega \cap \mathbb{S}^n_{++},\end{aligned} \tag{4}$$

where $f : \mathbb{R}^{n\times n} \to \mathbb{R}$ is a function of a matrix. A feasible point $\mathbf{X}^*$ of (4) is called a *strict local minimizer* of (4) if for any $\mathbf{Y} \in \mathbb{R}^{n\times n}$, $\epsilon > 0$ such that $\mathbf{X}^* + \mathbf{Y}$ is feasible with respect to the constraints of (4) and $\|\mathbf{Y}\|_F \in (0, \epsilon)$, then $f(\mathbf{X}^* + \mathbf{Y}) > f(\mathbf{X}^*)$.

3. Homotopy Approximate Newton Algorithm

This section presents a homotopy approximate Newton algorithm using warm starts (denoted by "HANA") intended to solve the ℓ_0-constrained inverse covariance learning Problem (3). We note that the set of optimal solutions for (3) can be unbounded, which may cause difficulty for any algorithm directly applied to (3). Thus, instead of directly solving (3), we propose solving a sequence of subproblems parameterized by $\lambda > 0$,

$$\begin{aligned}\min_{\mathbf{X}} \quad & f(\mathbf{X}) = f_0(\mathbf{X}) + \frac{\lambda}{2}\|\mathbf{X}\|_F^2 \\ \text{s.t.} \quad & \mathbf{X} \in \Omega \cap \mathbb{S}^n_{++},\end{aligned} \tag{5}$$

where we denote the Frobenius norm by $\|\mathbf{X}\|_F^2 = \sum_{i,j=1}^n |X_{ij}|^2$ and

$$f_0(\mathbf{X}) = \operatorname{tr}(\mathbf{SX}) - \log\det(\mathbf{X}). \tag{6}$$

We note that the gradient of the objective function in (5) is $\nabla f(\mathbf{X}) = \mathbf{S} - \mathbf{X}^{-1} + \lambda\mathbf{X}$ and that the Hessian is $\nabla^2 f(\mathbf{X}) = \mathbf{X}^{-1} \otimes \mathbf{X}^{-1} + \lambda\mathbf{I}_{n^2}$.

The objective of (5) is strongly convex, and the optimal solution set of (5) is bounded; we will explicitly prove this boundedness property later, but the strong convexity is obvious. The intention of the ℓ_2 regularizer is to improve the conditioning of the subproblems, possibly allowing for faster identification of local minima of (5). By driving λ to zero in (5), an optimal solution for (3), the true problem of interest, can be obtained. Thus, to yield a practical algorithm, we consider a sequence of strictly deceasing values $\{\lambda_\ell\}$; in the ℓ-th iteration of our algorithm, we will seek an approximate minimizer to (5) with $\lambda = \lambda_\ell$. On the $(\ell + 1)$-th iteration, we warm start the solution of the $(\ell + 1)$-th subproblem with the solution to the ℓ-th subproblem as an initial point.

3.1. First-Order Stationarity Conditions

We now characterize some properties of (5), in particular concerning the boundedness of the set of optimal solutions and optimality conditions for (3) and (5). In what follows, we assume that the ℓ_2-regularization parameter for (5) satisfies $\lambda > 0$. We define the projection operator

$$P_\Omega(\mathbf{X}) \triangleq \arg\min_{\mathbf{Y}\in\Omega} \|\mathbf{X} - \mathbf{Y}\|_F.$$

Remark 1. Because the cardinality constraint Ω is a nonconvex set, the operator P_Ω is generally set-valued; that is, projections are not unique. It has been shown that a point in $P_\Omega(\mathbf{X})$ can be quickly determined (Hager et al. 2016). If $\mathbf{X} \in \mathbb{R}^{n\times n}$, then first we set $X_{i,j} = 0$ for every $(i,j) \in \Im$. A member of $P_\Omega(\mathbf{X})$ is obtained by sorting the absolute values of the magnitudes of the n^2 entries in $\mathbf{X}$ and setting all but the κ largest values in $\mathbf{X}$ to zero, where ties are broken to ensure that the projected matrix is symmetric. We note that as a result of sorting n^2 matrix entries, the complexity of the projection operation is $\mathcal{O}(n^2\log(n))$.

First, we show in the following theorem that we can safely reformulate (5) in such a way that the feasible set is compact.

Theorem 2. *Consider the problem*

$$\begin{aligned}\min_{\mathbf{X}} \quad & f(\mathbf{X}) \\ \text{s.t.} \quad & \mathbf{X} \in \Omega \cap \mathbb{S}^n_{++} \\ & \underline{\sigma}\mathbf{I}_n \preceq \mathbf{X} \preceq \overline{\sigma}\mathbf{I}_n,\end{aligned} \tag{7}$$

where we have fixed constants

$$\overline{\sigma} = \frac{n^2 + \sqrt{n^4 + 2\lambda n f(t\mathbf{I}_n) - 2\lambda n^2}}{\lambda},$$
$$\underline{\sigma} = \exp\left(-f(t\mathbf{I}_n) + \frac{n\lambda}{2}\exp\left(-2f(t\mathbf{I}_n)\right)\overline{\sigma}^{2-2n}\right)\overline{\sigma}^{1-n},$$

and $t = \frac{-\mathrm{tr}(\mathbf{S}) + \sqrt{\mathrm{tr}(\mathbf{S})^2 + 4n^2\lambda}}{2n\lambda}$. *Then, the sets of global minima of* (5) *and* (7) *are nonempty and equivalent.*

Proof. Suppose $\mathbf{X}^* \in \mathbb{R}^{n\times n}$ is an optimal solution of (5) and $0 < \sigma_{\min} = \sigma_1 \le \ldots \le \sigma_n = \sigma_{\max}$ are the eigenvalues of $\mathbf{X}^*$. We seek lower and upper bounds for σ_1 and σ_n, respectively.

Consider $\overline{\mathbf{X}} = t\mathbf{I}_n$ for an arbitrary fixed scalar $t > 0$. Because $\kappa \ge n$, $\overline{\mathbf{X}}$ is feasible with respect to the constraints of (5). We have

$$\begin{aligned} f(\overline{\mathbf{X}}) \ge f(\mathbf{X}^*) &= \mathrm{tr}(\mathbf{S}\mathbf{X}^*) - \log\det(\mathbf{X}^*) + \frac{\lambda}{2}\|\mathbf{X}^*\|_F^2 \\ &\ge -\log\det(\mathbf{X}^*) + \frac{\lambda}{2}\sum_{i=1}^n X_{ii}^{*2} \\ &\ge -n\log(\sigma_n) + \frac{\lambda}{2n}\left(\sum_{i=1}^n X_{ii}^*\right)^2 \\ &\ge -n\log(\sigma_n) + \frac{\lambda}{2n}\sigma_n^2. \end{aligned} \tag{8}$$

The second inequality is because $\mathbf{S} \succeq \mathbf{0}$ and $\mathbf{X}^* \succ \mathbf{0}$ and, thus, $\mathrm{tr}(\mathbf{S}\mathbf{X}^*) \ge 0$. The third inequality uses the Cauchy-Schwarz inequality, and the last inequality holds because $\sum_{i=1}^n X_{ii}^* = \sum_{i=1}^n \sigma_i$ and $\sigma_i > 0$. Combining (8) with the fact that $\log(\sigma_n) \le \sigma_n - 1$, we get

$$f(\overline{\mathbf{X}}) \ge -n(\sigma_n - 1) + \frac{\lambda}{2n}\sigma_n^2. \tag{9}$$

The right-hand side of (9) tends to infinity as $\sigma_n \to \infty$; because $f(\overline{\mathbf{X}})$ is finite for the fixed value of t, we have that σ_n is bounded from above. Specifically, σ_n cannot exceed the larger of the two solutions to the quadratic equation $\frac{\lambda}{2n}\sigma_n^2 - n(\sigma_n - 1) - f(\overline{\mathbf{X}}) = 0$. That is,

$$\sigma_n \le \overline{\sigma} \triangleq \frac{n^2 + \sqrt{n^4 + 2\lambda n f(\overline{\mathbf{X}}) - 2\lambda n^2}}{\lambda} < \infty. \tag{10}$$

Likewise, we obtain a lower bound for σ_1 as

$$\begin{aligned} f(\overline{\mathbf{X}}) \ge f(\mathbf{X}^*) &\ge -\log\det(\mathbf{X}^*) \\ &\ge -\log(\sigma_1) - (n-1)\log(\overline{\sigma}), \end{aligned}$$

from which we conclude

$$\sigma_1 \ge \exp\left(-f(\overline{\mathbf{X}})\right)\overline{\sigma}^{1-n}. \tag{11}$$

We now show that the lower bound in (11) can be tightened further. Starting from an intermediate step in the derivation of (8), we have

$$\begin{aligned} f(\overline{\mathbf{X}}) &\ge -\log\det(\mathbf{X}^*) + \frac{\lambda}{2n}\left(\sum_{i=1}^n \sigma_i\right)^2 \\ &\ge -\log(\sigma_1) - (n-1)\log(\overline{\sigma}) + \frac{n\lambda}{2}\sigma_1^2 \\ &\ge -\log(\sigma_1) - (n-1)\log(\overline{\sigma}) + \frac{n\lambda}{2}\exp\left(-2f(\overline{\mathbf{X}})\right)\overline{\sigma}^{2-2n}. \end{aligned}$$

Thus,

$$\sigma_1 \ge \underline{\sigma} \triangleq \exp\left(-f(\overline{\mathbf{X}}) + \frac{n\lambda}{2}\exp\left(-2f(\overline{\mathbf{X}})\right)\overline{\sigma}^{2-2n}\right)\overline{\sigma}^{1-n} > 0. \tag{12}$$

From Equations (10) and (12), in order to tighten the bounds $\underline{\sigma}$ and $\overline{\sigma}$, we should choose t so that $f(\overline{\mathbf{X}}) = f(t\mathbf{I}_n)$ is minimized. Observe that $f(t\mathbf{I}_n) = \mathrm{tr}(\mathbf{S})t - n\log(t) + \frac{n\lambda}{2}t^2$ is a strongly convex one-dimensional function in t for $t > 0$. Thus, $f(t\mathbf{I}_n)$ is minimized when the necessary optimality condition $\frac{d}{dt}f(t\mathbf{I}_n) = 0$ is satisfied. We, therefore, choose

$$t = \frac{-\mathrm{tr}(\mathbf{S}) + \sqrt{\mathrm{tr}(\mathbf{S})^2 + 4n^2\lambda}}{2n\lambda},$$

which completes the proof the equivalent of (5) and (7).

The nonemptiness follows from the compactness of the feasible set of (7). □

Observe that as $\lambda \to 0$, both $\underline{\sigma} \to 0$ and $\overline{\sigma} \to \infty$. That is, Problem (7) is reduced to Problem (3). This illustrates why we generally cannot guarantee that the original Problem (3) has at least one global minimizer; such a guarantee would require an additional boundedness assumption, as will be stated in Theorem 6.

We now introduce in Theorem 3 necessary optimality conditions for Problem (5). For any positive definite matrix $\mathbf{H} \in \mathbb{S}_{++}^n$ and a scalar $t > 0$, we define

$$\begin{aligned} h_{t,\mathbf{H},\mathbf{X}}(\mathbf{Z}) &= \mathrm{tr}\left(\nabla f(\mathbf{X})(\mathbf{Z} - \mathbf{X})\right) + \frac{1}{2t}\|\mathbf{Z} - \mathbf{X}\|_{\mathbf{H}}^2, \\ \mathcal{S}_{t,\mathbf{H}}(\mathbf{X}) &= \arg\min_{\mathbf{Z}\in\Omega} h_{t,\mathbf{H},\mathbf{X}}(\mathbf{Z}) \\ &= \arg\min_{\mathbf{Z}\in\Omega} \|\mathbf{Z} - \left(\mathbf{X} - t\mathbf{H}\nabla f(\mathbf{X})\right)\|_{\mathbf{H}^{-1}}^2, \end{aligned} \tag{14}$$

where $\|\mathbf{X}\|_{\mathbf{H}} = \sqrt{\mathbf{X}^{\mathrm{T}}\mathbf{H}\mathbf{X}}$.

Theorem 3. *Suppose that* $\mathbf{X}^*$ *is an optimal solution of* (5) *and* $\mathbf{H}$ *is positive definite. Then there exists* $T > 0$ *such that*

$$\mathbf{X}^* = \mathcal{S}_{t,\mathbf{H}}(\mathbf{X}^*) \quad \textit{for all} \quad t \in (0, T). \tag{15}$$

Proof. Let $\overline{\mathbf{X}}$ be a feasible solution for (5). Define the level set $\mathcal{L}$ as

$$\mathcal{L} = \{\mathbf{X} \in \Omega \cap \mathbb{S}^n_{++} : f(\mathbf{X}) \le f(\overline{\mathbf{X}})\}.$$

We note that $f(\mathbf{X}^*) \le f(\overline{\mathbf{X}})$; thus, $\mathbf{X}^* \in \mathcal{L}$. We proved in Theorem 2 that $\mathcal{L} \subset \{\mathbf{X} \in \mathbb{R}^{n\times n} : \underline{\sigma}\mathbf{I}_n \le \mathbf{X} \le \overline{\sigma}\mathbf{I}_n\}$, where $\underline{\sigma}$ and $\overline{\sigma}$ are particular constants satisfying $0 < \underline{\sigma} < \overline{\sigma} < \infty$. Because $\nabla^2 f(\mathbf{X}) = \mathbf{X}^{-1} \otimes \mathbf{X}^{-1} + \lambda \mathbf{I}_{n^2}$, it follows that $\nabla^2 f(\mathbf{X}) \le (\underline{\sigma}^{-2} + \lambda)\mathbf{I}_{n^2}$ for every $\mathbf{X} \in \mathcal{L}$. Hence, we have

$$f(\mathbf{X}) \le f(\mathbf{X}^*) + \operatorname{tr}\big(\nabla f(\mathbf{X}^*)(\mathbf{X} - \mathbf{X}^*)\big) + \frac{1}{2L_f}\|\mathbf{X} - \mathbf{X}^*\|_F^2, \tag{16}$$

for all $\mathbf{X} \in \mathcal{L}$, where $L_f = (\underline{\sigma}^{-2} + \lambda)^{-1}$. We will show that (15) holds for $T = \sigma_1(\mathbf{H})L_f$.

First, we demonstrate that $\mathbf{X}^* \in \mathcal{S}_{t,\mathbf{H}}(\mathbf{X}^*)$ for any $t \in (0, T)$. Notice that this will not prove the theorem, as this claim is a statement concerning the existence of $\mathbf{X}^*$. We will subsequently prove the uniqueness of $\mathbf{X}^*$, which will thus prove the theorem. From (16) and because $\frac{1}{t}\mathbf{H} > \frac{1}{L_f}\mathbf{I}_n$ for all $t \in (0, T)$, we have for all $\mathbf{X} \in \mathcal{L}$

$$f(\mathbf{X}^*) \le f(\mathbf{X}) \le f(\mathbf{X}^*) + \operatorname{tr}\big(\nabla f(\mathbf{X}^*)(\mathbf{X} - \mathbf{X}^*)\big) + \frac{1}{2t}\|\mathbf{X} - \mathbf{X}^*\|_{\mathbf{H}}^2.$$

It follows that $h_{t,\mathbf{H},\mathbf{X}^*}(\mathbf{X}) = \operatorname{tr}(\nabla f(\mathbf{X}^*)(\mathbf{X} - \mathbf{X}^*)) + \frac{1}{2t}\|\mathbf{X} - \mathbf{X}^*\|_{\mathbf{H}}^2 \ge 0$ for all $\mathbf{X} \in \mathcal{L}$. Thus $\mathbf{X}^* \in \mathcal{S}_{t,\mathbf{H}}(\mathbf{X}^*)$ because $h_{t,\mathbf{H},\mathbf{X}^*}(\mathbf{X}^*) = 0$, proving the claim.

Now, to prove uniqueness, suppose toward contradiction that there exists $\hat{\mathbf{X}} \in \mathcal{S}_{t,\mathbf{H}}(\mathbf{X}^*)$ for some $0 < t < T$ but that $\hat{\mathbf{X}} \ne \mathbf{X}^*$. Noting that by definition $h_{t,\mathbf{H},\mathbf{X}^*}(\mathbf{X}^*) = 0$, we must have $h_{t,\mathbf{H},\mathbf{X}^*}(\hat{\mathbf{X}}) = 0$, and so

$$\operatorname{tr}\big(\nabla f(\mathbf{X}^*)(\hat{\mathbf{X}} - \mathbf{X}^*)\big) = -\frac{1}{2t}\big\|\hat{\mathbf{X}} - \mathbf{X}^*\big\|_{\mathbf{H}}^2. \tag{17}$$

Combining (16) and (17), we have

$$\begin{aligned} f(\hat{\mathbf{X}}) &\le f(\mathbf{X}^*) + \operatorname{tr}\Big(\nabla f(\mathbf{X}^*)\big(\hat{\mathbf{X}} - \mathbf{X}^*\big)\Big) + \frac{1}{2L_f}\big\|\hat{\mathbf{X}} - \mathbf{X}^*\big\|_F^2 \\ &= f(\mathbf{X}^*) - \frac{1}{2}\big(\hat{\mathbf{X}} - \mathbf{X}^*\big)^T\left(\frac{\mathbf{H}}{t} - \frac{\mathbf{I}_n}{L_f}\right)\big(\hat{\mathbf{X}} - \mathbf{X}^*\big) \\ &< f(\mathbf{X}^*), \end{aligned} \tag{18}$$

which contradicts the global optimality of $\mathbf{X}^*$ for (5). □

For a general positive definite matrix $\mathbf{H}$, solving the problem defining $\mathcal{S}_{t,\mathbf{H}}(\mathbf{X})$ in (14) is generally intractable because of the cardinality constraint. As a result, verifying the optimality condition (15) is computationally expensive. The following lemma offers a simpler condition for (5), which is an immediate consequence of Theorem 3 when $\mathbf{H} = \mathbf{I}_n$. A similar result was established in theorem 3.2 of Lu (2015) but for a different closed constraint set.

Lemma 2. *Suppose that $\mathbf{X}^*$ is an optimal solution of (5). Then there exists $T > 0$ such that*

$$\mathbf{X}^* = P_\Omega\big(\mathbf{X}^* - t\nabla f(\mathbf{X}^*)\big) \quad \text{for all} \quad t \in (0, T). \tag{19}$$

The optimality condition for the original Problem (3) can be stated as follows.

Lemma 3. *Suppose that the level set $\mathcal{L}_0 = \{\mathbf{X} \in \Omega \cap \mathbb{S}^n_{++} : f_0(\mathbf{X}) \le f_0(\mathbf{X}^{(0)})\}$ for Problem (3) for some $\mathbf{X}^{(0)} \in \Omega \cap \mathbb{S}^n_{++}$ is a compact set. If $\mathbf{X}^*$ is an optimal solution of (3), then there exists $T > 0$ such that*

$$\mathbf{X}^* = P_\Omega\big(\mathbf{X}^* - t\nabla f_0(\mathbf{X}^*)\big) \quad \text{for all} \quad t \in (0, T). \tag{20}$$

Proof. First we argue that there exist $0 < \underline{\sigma} < \overline{\sigma} < \infty$ such that $\underline{\sigma}\mathbf{I}_n \prec \mathbf{X} \prec \overline{\sigma}\mathbf{I}_n$ for any $\mathbf{X} \in \mathcal{L}_0$. Because $\mathcal{L}_0$ is bounded, there exists $M > 0$ such that $\|\mathbf{X}\|_F \le M$ for all $\mathbf{X} \in \mathcal{L}_0$. For any $\mathbf{X} \in \mathcal{L}_0$, we have

$$M^2 \ge \|\mathbf{X}\|_F^2 \ge \sum_{i=1}^n X_{ii}^2 \ge \frac{1}{n}\left(\sum_{i=1}^n X_{ii}\right)^2 \ge \frac{\sigma_{\max}^2(\mathbf{X})}{n}.$$

We, therefore, have $\sigma_{\max}(\mathbf{X}) \le M\sqrt{n} = \overline{\sigma}$. To derive a lower bound, we observe that

$$\begin{aligned} f_0(\mathbf{X}^{(0)}) &\ge -\log\det(\mathbf{X}) \\ &\ge -\log(\sigma_{\min}(\mathbf{X})) - (n-1)\log(\overline{\sigma}). \end{aligned}$$

This implies that $\sigma_{\min}(\mathbf{X}) \ge \underline{\sigma} = \exp(-f_0(\mathbf{X}^{(0)}))\overline{\sigma}^{1-n}$.

We have $\nabla^2 f_0(\mathbf{X}) = \mathbf{X}^{-1} \otimes \mathbf{X}^{-1}$. Hence, it follows that $\nabla^2 f_0(\mathbf{X}) \le \underline{\sigma}^{-2}\mathbf{I}_{n^2}$ for every $\mathbf{X} \in \mathcal{L}_0$. Therefore, a Lipschitz constant for f_0 over $\mathcal{L}_0$ is $L_{f_0} = \underline{\sigma}^2$. Now following the same argument for the proof of Theorem 3 by using $\mathbf{H} = \mathbf{I}_n$, we complete the proof. □

Our subsequent convergence analysis will rely on Lemmas 2 and 3. We will now state and prove a technical proposition regarding how the operator $\mathcal{S}_{t,\mathbf{I}_n}$ acts on convergent sequences. We will make use of this result in the proof of Theorem 4 for convergence to a stationary point.

Proposition 1. *Suppose $\{\mathbf{X}^k\}$ is a sequence in $\mathbb{R}^{n\times n}$ converging to $\mathbf{X}^*$, and suppose there exists a corresponding sequence $\{\mathbf{Y}^k\}$ such that $\mathbf{Y}^k \in \mathcal{S}_{t,\mathbf{I}_n}(\mathbf{X}^k)$ for each k. If there exists $\mathbf{Y}^*$ such that $\mathbf{Y}^k \to \mathbf{Y}^*$, then $\mathbf{Y}^* \in \mathcal{S}_{t,\mathbf{I}_n}(\mathbf{X}^*)$ for $t > 0$.*

Proof. Because $\mathbf{Y}^k \in \Omega$ for each k and Ω is closed, we conclude that $\mathbf{Y}^* \in \Omega$. Hence, we obtain from the definition of $\mathcal{S}_{t,\mathbf{I}_n}(\mathbf{X}^*)$ that

$$\begin{aligned} &\|\mathbf{Y}^* - (\mathbf{X}^* - t\nabla f(\mathbf{X}^*))\|_F \\ &\quad \ge \min_{\mathbf{Y}\in\Omega}\|\mathbf{Y} - (\mathbf{X}^* - t\nabla f(\mathbf{X}^*))\|_F. \end{aligned} \tag{21}$$

If inequality (21) holds as an equality, then we have proven the proposition. Toward a contradiction, suppose (21) holds as a strict inequality. Then, for all k,

$$\begin{aligned}\|\mathbf{Y}^* - (\mathbf{X}^* - t\nabla f(\mathbf{X}^*))\|_F &> \min_{\mathbf{Y}\in\Omega} \|\mathbf{Y} - (\mathbf{X}^* - t\nabla f(\mathbf{X}^*))\|_F \\ &\geq \Big|\min_{\mathbf{Y}\in\Omega} \|\mathbf{Y} - (\mathbf{X}^k - t\nabla f(\mathbf{X}^k))\|_F - \|(\mathbf{X}^* - t\nabla f(\mathbf{X}^*)) \\ &\quad - (\mathbf{X}^k - t\nabla f(\mathbf{X}^k))\|_F\Big| \\ &= \Big|\|\mathbf{Y}^k - (\mathbf{X}^k - t\nabla f(\mathbf{X}^k))\|_F - \|(\mathbf{X}^* - t\nabla f(\mathbf{X}^*)) \\ &\quad - (\mathbf{X}^k - t\nabla f(\mathbf{X}^k))\|_F\Big|, \end{aligned} \tag{22}$$

where the second inequality follows from the triangle inequality and the equality follows from the definition of $\mathbf{Y}^k$. As $k \to \infty$, (22) implies

$$\|\mathbf{Y}^* - (\mathbf{X}^* - t\nabla f(\mathbf{X}^*))\|_F > \|\mathbf{Y}^* - (\mathbf{X}^* - t\nabla f(\mathbf{X}^*))\|_F,$$

a contradiction. □

3.2. Algorithm Description

In this section, we present an algorithm to solve Problem (5). Our strategy for solving the constrained optimization problem is to embed an approximate Newton method into a generalized gradient projection scheme. In this algorithm, feasibility with respect to the sparsity constraint Ω is handled via projection, whereas the symmetric positive definiteness of the iterations (i.e., feasibility with respect to $\{\mathbf{X} \succ \mathbf{0}\}$) is ensured through a line search procedure.

A particular issue in the application of a projected gradient method to (5) is that the constraint set $\Omega \cap \mathbb{S}^n_{++}$ of (5) is nonconvex and not closed. This is a completely different situation from problems in the conventional scope of projected gradient methods, which require the feasible set to be convex and closed (Birgin et al. 2000). The issue of nonconvex constraint sets in a projected gradient framework has been successfully navigated in some special cases where the projection onto the constraint set is well defined (even if generally nonunique) and reasonably inexpensive in terms of computational cost (Hager et al. 2016, Xu et al. 2016). However, we cannot directly apply these methods for our problem (5) because of the nonclosedness of the positive definite constraint. We avoid dealing with the positive definite constraint in the projection step; for the same reason, the optimality condition in Lemma 2 is related only to the projection onto Ω, not onto $\Omega \cap \mathbb{S}^n_{++}$.

In our algorithm, we alternatively employ the negative gradient direction $-\nabla f(\mathbf{X})$ and an approximate Newton direction $\mathbf{D}$. We compute Newton directions only in a lower-dimensional subspace of free variables. Given $\mathbf{X}^k$ at the k-th iteration, the next iteration is updated by using either the gradient direction

$$\mathbf{X}^{k+1} \leftarrow P_\Omega(\mathbf{X}^k - \alpha_k \nabla f(\mathbf{X}^k)) \tag{23}$$

or the approximate Newton direction

$$\mathbf{X}^{k+1} \leftarrow P_\Omega(\mathbf{X}^k + \alpha_k \mathbf{D}^k). \tag{24}$$

We will now discuss how to compute the approximate Newton direction $\mathbf{D}^k$, how to determine stepsize α_k, and how to select the type of update steps to be computed in the k-th iteration.

3.2.1. Computing the Approximate Newton Direction.

To accelerate convergence and make use of the fact that the optimal solution is often sparse in many applications, that is, $\kappa \ll n^2$, we incorporate projected approximate Newton steps into our method by using second-order information in a reduced subspace. At each point $\mathbf{X}^k$, a local quadratic model of the objective function can be minimized subject to the constraint that all but a *working set* $\mathcal{W}_k$ of the matrix entries will not be fixed to zero in the minimization. For any working set $\mathcal{W}_k \subseteq \{1,\dots,n\} \times \{1,\dots,n\}$ and a matrix $\mathbf{Y}$, we define $[\mathbf{Y}]_{\mathcal{W}_k} \in \mathbb{R}^{n\times n}$ coordinate-wise via

$$\left(\mathbf{Y}_{\mathcal{W}_k}\right)_{(i,j)} = \begin{cases} Y_{ij} & \text{if } (i,j) \in \mathcal{W}_k \\ 0 & \text{otherwise.} \end{cases}$$

Then, given a current iteration $\mathbf{X}^k$, gradient $\nabla f(\mathbf{X}^k)$, and Hessian $\nabla^2 f(\mathbf{X}^k)$, we solve the following reduced quadratic problem to obtain a Newton direction $[\mathbf{D}]_{\mathcal{W}}k$:

$$\begin{aligned}\min_{[\mathbf{D}]_{\mathcal{W}_k}\in\mathbb{R}^{n\times n}} &\operatorname{tr}\left(\nabla f(\mathbf{X}^k)[\mathbf{D}]_{\mathcal{W}_k}\right) \\ &+ \frac{1}{2}\operatorname{vec}\left([\mathbf{D}]_{\mathcal{W}_k}\right)^{\mathrm{T}} \nabla^2 f(\mathbf{X}^k) \operatorname{vec}\left([\mathbf{D}]_{\mathcal{W}_k}\right). \end{aligned} \tag{25}$$

To efficiently obtain an approximate solution of (25), we use a conjugate gradient method proposed in (Olsen et al. 2012). This method takes advantage of the particular Kronecker product structure of $\nabla^2 f(\mathbf{X}^k)$, as well as the observation that $|\mathcal{W}_k| \ll n^2$, because we typically search for a very sparse solution in an abundance of applications.

3.2.2. Identifying the Working Set.

The working set includes the coordinates associated with nonzeros of $\mathbf{X}^k$, that is, $\{(i,j) : X^k_{i,j} \neq 0\} \subset \mathcal{W}_k$. The working set also includes the coordinates that violate the optimality conditions given in Lemma 2. We proved in Lemma 2 that if $\mathbf{X}^k$ is an optimal solution of (5), then $P_\Omega(\mathbf{X}^k - \alpha_k \nabla f(\mathbf{X}^k)) - \mathbf{X}^k = \mathbf{0}$ for a small α_k. Thus, we include in $\mathcal{W}_k$ the indices (i,j) that correspond to $(P_\Omega(\mathbf{X}^k - \alpha_k \nabla f(\mathbf{X}^k)))_{i,j} - X^k_{i,j} \neq 0$. More specifically, if $\mathbf{X}^k$ is updated via (23), then $\mathcal{W}_k$ is defined as

$$\mathcal{W}_k = \left\{(i,j) : X^k_{i,j} \neq 0\right\} \cup \left\{(i,j) : |X^k_{i,j} - X^{k-1}_{i,j}| \geq \epsilon\right\} \tag{26}$$

for some small number $\epsilon > 0$. Otherwise, if $\mathbf{X}^k$ is updated via the approximate Newton step (24), then the working set takes the form

$$\mathcal{W}_k = \left\{(i,j) : X^k_{i,j} \neq 0\right\} \cup \left\{(i,j) : \left|Y^k_{i,j} - X^k_{i,j}\right| \geq \epsilon\right\}, \quad (27)$$

where $\mathbf{Y}^k \in P_\Omega(\mathbf{X}^k - \alpha_k \nabla f(\mathbf{X}^k))$.

We can see that $\{(i,j) : X^k_{i,j} \neq 0\} \subseteq \mathcal{C} \triangleq \{(i,j) : X^k_{i,j} \neq 0 \text{ or } Y^k_{i,j} \neq 0\}$ and $\{(i,j) : |Y^k_{i,j} - X^k_{i,j}| \geq \epsilon\} \subseteq \mathcal{C}$. Hence, we have $\mathcal{W}_k \subseteq \mathcal{C}$. Because $\mathcal{C}$ has at most cardinality 2κ, it follows that the dimension of the subspace defined by the working space is at most 2κ; in other words, $|\mathcal{W}_k| \leq 2\kappa$, which is small in many applications.

3.2.3. Choosing the Stepsize. On each gradient search step (23), given a feasible point $\mathbf{X}^k$ and an initial stepsize α_k, we backtrack along the projection arc defined by $P_\Omega(\mathbf{X}^k - \alpha_k \nabla f(\mathbf{X}^k))$ to determine a trial point. We initialize the stepsize $\alpha_k > 0$ with a Barzilai-Borwein (BB) step (Barzilai and Borwein 1988):

$$\alpha^{BB}_k = \frac{\operatorname{tr}\left(\left(\nabla f(\mathbf{X}^k) - \nabla f(\mathbf{X}^{k-1})\right)^{\mathsf{T}}\left(\mathbf{X}^k - \mathbf{X}^{k-1}\right)\right)}{\|\mathbf{X}^k - \mathbf{X}^{k-1}\|^2_F}.$$

It has been empirically observed that BB-stepsizes perform much better when they are embedded in a nonmonotone line search. This motivates us to investigate a scheme based on the Grippo-Lampariello-Lucidi stepsize rule (Grippo et al. 1986). Let $f^k_{\max}$ denote the largest of the M most recent function values:

$$f^k_{\max} = \max\{f(\mathbf{X}^{k-j}) : 0 \leq j < \min(k, M)\}.$$

We use this quantity $f^k_{\max}$ to define a sufficient decrease condition. In particular, we terminate the backtracking procedure for computing $\mathbf{X}^{k+1} = P_\Omega(\mathbf{X}^k - \alpha_k \nabla f(\mathbf{X}^k))$ by gradually reducing α_k when two conditions (C1) and (C2) are satisfied:

(C1) The objective function value $f(\mathbf{X}^{k+1})$ is sufficiently decreased:

$$f(\mathbf{X}^{k+1}) \leq f^k_{\max} - \frac{\delta}{2}\left\|\mathbf{X}^{k+1} - \mathbf{X}^k\right\|^2_F,$$

for some positive parameter $\delta > 0$.

(C2) The iteration $\mathbf{X}^{k+1}$ is positive definite.

The same line-search is used for the Newton direction update (24), where we initialize $\alpha_k = 1$.

3.2.4. Selecting the Update Steps. If we iteratively searched for descent directions only within a sequence of working set subspaces, then there is an intuitive possibility that a projected approximate Newton algorithm will converge to a low-quality local minimizer. Occasionally performing a line search in the projected gradient direction in the full space seems to help a projected approximate Newton algorithm identify a good working set subspace in practice. Moreover, the analysis of the projected gradient directions allows us to prove global convergence results of our method. Thus, the general logic of our algorithm is that, for fixed $q \in \mathbb{N}$, after every $q-1$ iterations of the Newton steps we perform a single iteration of the gradient projection step.

A detailed description of our algorithm `ANA` for solving (5) is given in Algorithm 1.

Algorithm 1 (Approximate Newton Algorithm `ANA`$(\mathbf{X}^0, \Omega, \lambda)$)

Choose parameters $\sigma \in (0,1), \delta > 0, \epsilon > 0, \alpha^{Newton}_{\min} > 0$, $[\alpha_{\min}, \alpha_{\max}] \subset (0,\infty)$, integers p, q such that $0 < p < q$, integer $M > 0$, *grad_step* = **true**.
Set $k = 0$.
while some stopping criteria not satisfied **do**
  Step 1: Approximate Newton step
  if $(mod(k,q) \neq p)$ **then**
    if (*grad_step* = **true**) **then**
      $\mathcal{W}_k \leftarrow$ Eq. (26)
    else
      $\mathcal{W}_k \leftarrow$ Eq. (27)
    Approximately solve for $\mathbf{D}^k$ from (25)
    Choose $\alpha_k = \alpha_{k,0} = 1$
    $j \leftarrow 0$, *ls* $\leftarrow$ **true**
    while (*ls* = **true**) & $(\alpha_k \geq \alpha^{Newton}_{\min})$ **do**
      $\alpha_k \leftarrow \sigma^j \alpha_{k,0}$
      $\mathbf{X}^{k+1}_{trial} \leftarrow P_\Omega(\mathbf{X}^k + \alpha_k \mathbf{D}^k)$
      if $(f(\mathbf{X}^{k+1}) \leq f^k_{\max} - \frac{\delta}{2}\|\mathbf{X}^{k+1} - \mathbf{X}^k\|^2_F$ and $\mathbf{X}^{k+1} > \mathbf{0})$ **then**
        $\mathbf{X}^{k+1} \leftarrow \mathbf{X}^{k+1}_{trial}$
        ls $\leftarrow$ **false**
      else
        $j \leftarrow j + 1$
  Step 2: Gradient projection step
  grad_step $\leftarrow$ **false**
  if $(mod(k,q) = p)$ *or* $(\alpha_k < \alpha^{Newton}_{\min})$ **then**
    Choose $\alpha_{k,0} = \min(\alpha_{\max}, \max(\alpha_{\min}, \alpha^{BB}_k))$
    $j \leftarrow 0$, *ls* $\leftarrow$ **true**
    while (*ls* = **true**) **do**
      $\alpha_k \leftarrow \sigma^j \alpha_{k,0}$
      $\mathbf{X}^{k+1}_{trial} \leftarrow P_\Omega(\mathbf{X}^k - \alpha_k \nabla f(\mathbf{X}^k))$
      if $(f(\mathbf{X}^{k+1}) \leq f^k_{\max} - \frac{\delta}{2}\|\mathbf{X}^{k+1} - \mathbf{X}^k\|^2_F$ and $\mathbf{X}^{k+1} > \mathbf{0})$ **then**
        $\mathbf{X}^{k+1} \leftarrow \mathbf{X}^{k+1}_{trial}$
        ls $\leftarrow$ **false**
      else
        $j \leftarrow j + 1$
    grad_step $\leftarrow$ **true**
  Step 3: Iteration
  $k \leftarrow k + 1$.

Naively, on each iteration of the backtracking procedure, one would compute an eigenvalue decomposition of the trial step $\mathbf{X}^{k+1}$ in order to both compute the value of $\det(\mathbf{X}^{k+1})$ in the objective function and to determine the satisfaction of (C2). Computing a determinant is generally an expensive operation. Because we are explicitly searching for a positive definite matrix, however, it is enough to attempt a Cholesky factorization of $\mathbf{X}^{k+1}$, which can be done more efficiently. If the factorization fails (a complex number appears on the diagonal of the lower triangular matrix), then the current $\mathbf{X}^{k+1}$ is not positive definite and we can immediately retract along the projection arc. If the factorization is successful, then the determinant of the matrix $\mathbf{X}^{k+1}$ can be computed from the diagonal values of the lower triangular matrix. This sequence of attempted Cholesky factorizations is repeated until the stopping criteria is satisfied. Finite termination of the backtracking step is proven in Section 4.

Algorithm 2 (Homotopy Approximate Newton Algorithm Using Warm Starts—HANA)

Choose parameters $L, \kappa_0, \lambda_0, \lambda_L$ and an initial point $\mathbf{X}^{(0)}$.
Calculate homotopy sequence $\{(\kappa_\ell, \lambda_\ell) : \ell = 0, 1, \ldots, L\}$:
  $\kappa_0 \geq \kappa_1 \geq \cdots \geq \kappa_L = \kappa,\ \lambda_0 \geq \lambda_1 \geq \cdots \geq \lambda_L > 0.$
for $\ell = 0, 1, \ldots, L$ **do**
  $\Omega_\ell \leftarrow \{\mathbf{X} : \|\mathbf{X}\|_0 \leq \kappa_\ell,\ \ X_{i,j} = 0\ \ \forall (i,j) \in \Im\}$
  $\mathbf{X}^{(\ell)} \leftarrow P_{\Omega \cap \mathbb{S}^n_{++}}(\mathbf{X}^{(\ell)})$
  $\mathbf{X}^{(\ell+1)} \leftarrow \text{ANA}(\mathbf{X}^{(\ell)}), \Omega_\ell, \lambda_\ell)$

3.2.5. Homotopy Scheme. We recall that the difference between (5), which Algorithm 1 is intended to solve, and (3), the real problem of interest, lies in the ℓ_2 regularization term. The regularized Problem (5) is expected to be solved more efficiently than the original Problem (3) because the objective of (5) is strongly convex and its optimal solution set is bounded. A solution of (3) can be obtained by tracing the homotopy solution path in a series of decreasing values of parameter λ. We note that, as is typically the case in nonconvex optimization, the iteratations of Algorithm 1 need not converge to a global minimizer of (5) but only to a point satisfying necessary conditions for optimality. To improve solution quality, we initially expand the subspace of free variables by using larger values of κ. Thus, we propose an outer homotopy method for solving (3). Algorithm 2 solves a sequence of subproblems of the form (5) by calling Algorithm 1, attempting to find increasingly better-quality minimizers. Two decreasing sets of integer values $\kappa_0 \geq \kappa_1 \geq \ldots \geq \kappa_L = \kappa$ and real values $\lambda_0 \geq \lambda_1 \geq \ldots \geq \lambda_L > 0$ are selected before running the algorithm. We first (approximately) solve (5) using a sparsity parameter κ_0, regularization parameter λ_0, and some initial point $\mathbf{X}^{(0)}$. Then, we sequentially solve (5) to some tolerance for each pair $(\kappa_\ell, \lambda_\ell)$: $\ell = 0, 1, \ldots, L$, initializing each subproblem solve with the previous solution $\mathbf{X}^{(l-1)}$. Note that the final regularization parameter λ_L should be chosen quite small and $\kappa_L = \kappa$ in practice. Here, we denote $P_{\Omega \cap \mathbb{S}^n_{++}}(\mathbf{X})$ by the projection of matrix $\mathbf{X}$ onto $\Omega \cap \mathbb{S}^n_{++}$, which can be done by alternating projection.

4. Convergence Analysis

We will prove in this section that the iterations of Algorithm 1 for solving Problem (5) accumulate at a point satisfying the necessary optimality conditions (19). Similarly, the second algorithm HANA converges to a stationary point of (3). Any reference to an objective function f in this section is specifically referring to the objective of (5).

We state an assumption on the initial point:

Assumption 2. *The initial point $\mathbf{X}^0$ is feasible with respect to the constraints of* (5)*, that is,* $\mathbf{X}^0 \in \Omega \cap \mathbb{S}^n_{++}$.

In what follows, we suppose that Problem (5) satisfies Assumptions 1 and 2. We first demonstrate that the line search in either Step 1 (the approximate Newton step) or Step 2 (the gradient projection step) terminates in a finite number of iterations.

Lemma 4. *Let $\{\mathbf{X}^k\}$ denote the sequence of iterations generated by Algorithm 1 for solving Problem* (5)*. Suppose the line search starting in either line* 13 *or line* 26 *of Algorithm 1 is reached in the k-th iteration and suppose $\mathbf{X}^k$ is feasible with respect to the constraints of* (5)*. Then, the line search in Steps* 1 *and* 2 *of Algorithm 1 terminates in a finite number of iterations.*

Proof. One can easily see that the line search in Step 1 terminates in a finite number of iterations because we enforce the termination condition $\alpha_k < \alpha_{\min}^{Newton}$. We now prove the finite termination when the line search occurs in Step 2 (the gradient projection step). We assume that $\nabla f(\mathbf{X}^k) \neq \mathbf{0}$, because otherwise the algorithm stops at a local solution $\mathbf{X}^k$.

We will first show that there exists $T_p^k > 0$ such that every matrix in $P_\Omega(\mathbf{X}^k - t\nabla f(\mathbf{X}^k))$ is positive definite for all $t \in (0, T_p^k)$. For a nonzero symmetric matrix $\mathbf{U}$, we will denote the spectral norm of $\mathbf{U}$ by $\|\mathbf{U}\|_2 \triangleq \max(|\sigma_{\min}(\mathbf{U})|, |\sigma_{\max}(\mathbf{U})|) > 0$. Thus, for all $t \in (0, L_1)$, we have $\mathbf{X}^k + t\mathbf{U} > 0$, where we have defined $L_1 = \frac{\sigma_{\min}(\mathbf{X}^k)}{\|\mathbf{U}\|_2}$.

Given $\mathbf{X}^k$, let $\mathcal{I}_k$ and $\mathcal{A}_k$ denote, respectively, the inactive and active matrix indices of $\mathbf{X}^k$ with respect to the cardinality constraint:

$$\mathcal{I}_k \triangleq \left\{(i,j) : X^k_{i,j} \neq 0\right\},\quad \mathcal{A}_k \triangleq \left\{(i,j) : X^k_{i,j} = 0\right\}.$$

We additionally define $(i_x, j_x) \triangleq \arg\min_{(i,j)\in\mathcal{I}_k} |X^k_{i,j}|$, $(i_{g,n}, j_{g,n}) \triangleq \arg\max_{(i,j)\in\mathcal{I}_k} |(\nabla f(\mathbf{X}^k))_{i,j}|$, and $(i_{g,a}, j_{g,a}) \triangleq \arg\max_{(i,j)\in\mathcal{A}_k} |(\nabla f(\mathbf{X}^k))_{i,j}|$. We will demonstrate the

existence of the required $T_p^k > 0$ by analyzing two separate cases concerning the cardinality of the set of inactive indices, $|\mathcal{I}_k|$.

Case 1 ($|\mathcal{I}_k| = \kappa$): For every $(i,j) \in \mathcal{I}_k$ we have

$$\begin{aligned}|X_{i,j}^k - t\big(\nabla f(\mathbf{X}^k)\big)_{i,j}| &\geq |X_{i,j}^k| - t|\big(\nabla f(\mathbf{X}^k)\big)_{i,j}| \\ &\geq |X_{i_x,j_x}^k| - t|\big(\nabla f(\mathbf{X}^k)\big)_{i_{g,n},j_{g,n}}|.\end{aligned}$$

Defining $T_1^k = \frac{|X_{i_x,j_x}^k|}{|(\nabla f(\mathbf{X}^k))_{i_{g,n},j_{g,n}}| + |(\nabla f(\mathbf{X}^k))_{i_{g,a},j_{g,a}}|}$, we have from the above that for all $(i,j) \in \mathcal{I}_k$,

$$|X_{i,j}^k - t\big(\nabla f(\mathbf{X}^k)\big)_{i,j}| > t|\big(\nabla f(\mathbf{X}^k)\big)_{i_{g,a},j_{g,a}}| \geq t|\big(\nabla f(\mathbf{X}^k)\big)_{h,l}|, \tag{28}$$

for all $(h,l) \in \mathcal{A}_k$ and for all $t \in (0, T_1^k)$. Note that T_1^k is well defined because $|X_{i_x,j_x}^k|$ and $|(\nabla f(\mathbf{X}^k))_{i_{g,n},j_{g,n}}| + |(\nabla f(\mathbf{X}^k))_{i_{g,a},j_{g,a}}|$ are strictly positive. We can decompose

$$\begin{aligned}\mathbf{X}^k - t\nabla f(\mathbf{X}^k) &= \big[\mathbf{X}^k - t\nabla f(\mathbf{X}^k)\big]_{\mathcal{I}_k} + \big[\mathbf{X}^k - t\nabla f(\mathbf{X}^k)\big]_{\mathcal{A}_k} \\ &= \big[\mathbf{X}^k - t\nabla f(\mathbf{X}^k)\big]_{\mathcal{I}_k} - t\big[\nabla f(\mathbf{X}^k)\big]_{\mathcal{A}_k}.\end{aligned}$$

Combining this decomposition with Remark 1 and (28), we have that $[\mathbf{X}^k - t\nabla f(\mathbf{X}^k)]_{\mathcal{I}_k} \in P_\Omega(\mathbf{X}^k - t\nabla f(\mathbf{X}^k))$ for $t \in (0, T_1^k)$.

Now consider $\mathbf{U} = [\nabla f(\mathbf{X}^k)]_{\mathcal{I}_k}$ and define

$$T_2^k \triangleq \begin{cases} \frac{\sigma_1(\mathbf{X}^k)}{\|\mathbf{U}\|_2} & \text{if } \mathbf{U} \neq \mathbf{0} \\ 1 & \text{otherwise.}\end{cases}$$

Then, letting $T_p^k = \min(T_1^k, T_2^k)$, we have that $\mathbf{X}^k - t[\nabla f(\mathbf{X}^k)]_{\mathcal{I}_k} \in P_\Omega(\mathbf{X}^k - t\nabla f(\mathbf{X}^k))$ for all $t \in (0, T_p^k)$. Moreover, $\mathbf{X}^k - t[\nabla f(\mathbf{X}^k)]_{\mathcal{I}_k}$ is positive definite for all $t \in (0, T_p^k)$.

Case 2 ($|\mathcal{I}_k| < \kappa$): Denote by $\widehat{\mathcal{A}_k} \subseteq \mathcal{A}_k$ the set of indices $(i,j) \in \mathcal{A}_k$ corresponding to the $\kappa - |\mathcal{I}_k|$ largest absolute magnitude entries of $[\nabla f(\mathbf{X}^k)]_{\mathcal{A}_k}$. Denote

$$(i_{g,a'}, j_{g,a'}) \triangleq \arg\max_{(i,j) \in \mathcal{A}_k \setminus \widehat{\mathcal{A}_k}} \Big|\big(\nabla f(\mathbf{X}^k)\big)_{i,j}\Big|.$$

Define

$$T_3^k = \frac{|X_{i_x,j_x}^k|}{\Big|\big(\nabla f(\mathbf{X}^k)\big)_{i_{g,n},j_{g,n}}| + |\big(\nabla f(\mathbf{X}^k)\big)_{i_{g,a'},j_{g,a'}}|\Big|}.$$

From (28), it follows that

$$|X_{i,j}^k + t\big(\nabla f(\mathbf{X}^k)\big)_{i,j}| > t|\big(\nabla f(\mathbf{X}^k)\big)_{i_{g,a'},j_{g,a'}}| \geq t|\big(\nabla f(\mathbf{X}^k)\big)_{k,h}| \tag{29}$$

for all $(i,j) \in \mathcal{I}_k$, for all $(k,h) \in \mathcal{A}_k \setminus \widehat{\mathcal{A}_k}$, and for all $t \in (0, T_3^k)$. Similarly to Case 1, we decompose

$$\begin{aligned}\mathbf{X}^k - t\nabla f(\mathbf{X}^k) &= \big[\mathbf{X}^k - t\nabla f(\mathbf{X}^k)\big]_{\mathcal{I}_k} + \big[\mathbf{X}^k - t\nabla f(\mathbf{X}^k)\big]_{\widehat{\mathcal{A}_k}} \\ &\quad + \big[\mathbf{X}^k - t\nabla f(\mathbf{X}^k)\big]_{\mathcal{A}_k \setminus \widehat{\mathcal{A}_k}} \\ &= \Big[\mathbf{X}^k - t\nabla f(\mathbf{X}^k)^k\Big]_{\mathcal{I}_k} - t\big[\nabla f(\mathbf{X}^k)\big]_{\widehat{\mathcal{A}_k}} \\ &\quad - t\big[\nabla f(\mathbf{X}^k)\big]_{\mathcal{A}_k \setminus \widehat{\mathcal{A}_k}}.\end{aligned}$$

From this decomposition, Remark 1, and (29), we have that $[\mathbf{X}^k - t\nabla f(\mathbf{X}^k)]_{\mathcal{I}_k \cup \widehat{\mathcal{A}_k}} \in P_\Omega(\mathbf{X}^k - t\nabla f(\mathbf{X}^k))$ for all $t \in (0, T_3^k)$. Define $\mathbf{U} = [\nabla f(\mathbf{X}^k)]_{\mathcal{I}_k \cup \widehat{\mathcal{A}_k}}$. Analogous to the definition of T_2^k in Case 1, define

$$T_4^k \triangleq \begin{cases} \frac{\sigma_1(\mathbf{X}^k)}{\|\mathbf{U}\|_2} & \text{if } \mathbf{U} \neq \mathbf{0} \\ 1 & \text{otherwise.}\end{cases}$$

Let $T_p^k = \min(T_3^k, T_4^k)$. For all $t \in (0, T_p^k)$, we have that $\mathbf{X}^k - t[\nabla f(\mathbf{X}^k)]_{\mathcal{I}_k \cup \widehat{\mathcal{A}_k}} \in P_\Omega(\mathbf{X}^k - t\nabla f(\mathbf{X}^k))$. Moreover, $\mathbf{X}^k - t[\nabla f(\mathbf{X}^k)]_{\mathcal{I}_k \cup \widehat{\mathcal{A}_k}}$ is positive definite. Because every matrix in $P_\Omega(\mathbf{X}^k - t\nabla f(\mathbf{X}^k))$ can be expressed as $\mathbf{X}^k - t[\nabla f(\mathbf{X}^k)]_{\mathcal{I}_k \cup \widehat{\mathcal{A}_k}}$, we have proven the existence of the claimed T_p^k.

We now prove that there exists a positive number $T_\delta > 0$, independent of k, such that

$$f(\mathbf{X}^{k+1}) \leq f_{\max}^k - \frac{\delta}{2}\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2 \text{ provided } t \in (0, T_\delta), \tag{30}$$

where $\mathbf{X}^{k+1} \in P_\Omega(\mathbf{X}^k - t\nabla f(\mathbf{X}^k))$ and $\delta > 0$ is from Algorithm 1.

Suppose that $f(\mathbf{X}^k) \leq f_{\max}^{k-1} - \frac{\delta}{2}\|\mathbf{X}^k - \mathbf{X}^{k-1}\|^2$, where we have adopted the notation that $f_{\max}^{-1} = f(\mathbf{X}^0)$ and the convention[1] that $\mathbf{X}^{-1} = \mathbf{X}^0$. Define the level set

$$\mathcal{L} = \{\mathbf{X} \in \Omega \cap \mathbb{S}_{++}^n : f(\mathbf{X}) \leq f(\mathbf{X}^0)\}.$$

Note that $f(\mathbf{X}^k) \leq f_{\max}^{k-1} \leq f(\mathbf{X}^0)$ and so $\mathbf{X}^k \in \mathcal{L}$. Replacing $\mathbf{X}^*$ and $\overline{\mathbf{X}}$ by $\mathbf{X}$ and $\mathbf{X}^0$, respectively, in (10) and (12) yields $\mathcal{L} \subset \{\mathbf{X} \in \mathbb{R}^{n\times n} : \underline{\sigma}\mathbf{I}_n \preceq \mathbf{X} \preceq \overline{\sigma}\mathbf{I}_n\}$, where $\underline{\sigma}$ and $\overline{\sigma}$ are particular constants satisfying $0 < \underline{\sigma} < \overline{\sigma} < \infty$ from Theorem 2. Define

$$\mathcal{H} = \{\mathbf{X} \in \mathbb{R}^{n\times n} : 0.5\underline{\sigma}\mathbf{I}_n \preceq \mathbf{X} \preceq (\overline{\sigma} + 0.5\underline{\sigma})\mathbf{I}_n\}. \tag{31}$$

Obviously, we have $\mathcal{L} \subset \mathcal{H}$. Because $\mathbf{X}^{k+1}$ minimizes $h_{t,\mathbf{I}_n,\mathbf{X}^k}(\mathbf{Z})$, it follows that

$$\begin{aligned}h_{t,\mathbf{I}_n,\mathbf{X}^k}(\mathbf{X}^{k+1}) &= \mathrm{tr}\Big(\nabla f(\mathbf{X}^k)^\top(\mathbf{X}^{k+1} - \mathbf{X}^k)\Big) + \frac{1}{2t}\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2 \\ &\leq h_{t,\mathbf{I}_n,\mathbf{X}^k}(\mathbf{X}^k) = 0.\end{aligned}$$

We thus obtain

$$\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2 \leq 2t\mathrm{tr}\Big(\nabla f(\mathbf{X}^k)^\top(\mathbf{X}^k - \mathbf{X}^{k+1})\Big),$$

and it follows that

$$\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F \leq 2t\|\nabla f(\mathbf{X}^k)\|_F = 2t\|\mathbf{S} - (\mathbf{X}^k)^{-1} + \lambda\mathbf{X}^k\|_F.$$

Because $\mathbf{X}^k$ lies in the compact set $\mathcal{H}$, there exists a constant c, independent of $\mathbf{X}^k \in \mathcal{H}$, such that

$$\|\nabla f(\mathbf{X}^k)\|_F = \|\mathbf{S} - (\mathbf{X}^k)^{-1} + \lambda\mathbf{X}^k\|_F \leq c. \tag{32}$$

Let $\beta \in (0, \infty)$ satisfy $\beta c \leq 0.5\underline{\sigma}$. We claim that $\mathbf{X}^{k+1} \in \mathcal{H}$ provided that both $t \leq \beta$ and $\mathbf{X}^k \in \mathcal{L}$. Indeed, we have

$$\begin{aligned}\sigma_{\min}(\mathbf{X}^{k+1}) &\geq \sigma_{\min}(\mathbf{X}^k) - \sigma_{\max}(\mathbf{X}^{k+1} - \mathbf{X}^k)\\ &= \sigma_{\min}(\mathbf{X}^k) - \|\mathbf{X}^{k+1} - \mathbf{X}^k\|_2\\ &\geq \underline{\sigma} - \|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F\\ &\geq \underline{\sigma} - \beta c \geq 0.5\underline{\sigma}.\end{aligned}$$

Likewise, we have

$$\begin{aligned}\sigma_{\max}(\mathbf{X}^{k+1}) &\leq \sigma_{\max}(\mathbf{X}^k) + \sigma_{\max}(\mathbf{X}^{k+1} - \mathbf{X}^k)\\ &\leq \overline{\sigma} + \|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F \leq \overline{\sigma} + 0.5\underline{\sigma}.\end{aligned}$$

Hence, we conclude that $\mathbf{X}^{k+1} \in \mathcal{H}$. Because the eigenvalues of any matrix in $\mathcal{H}$ are bounded, f is Lipschitz continuously differentiable on $\mathcal{H}$. Denote L_f by the Lipschitz constant for f on $\mathcal{H}$. Because of the convexity of $\mathcal{H}$, we see that

$$\begin{aligned}f(\mathbf{X}^{k+1}) &\leq f(\mathbf{X}^k) + \operatorname{tr}\left(\nabla f(\mathbf{X}^k)^{\mathsf{T}}(\mathbf{X}^{k+1} - \mathbf{X}^k)\right)\\ &\quad + \frac{1}{2L_f}\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2\\ &\leq f(\mathbf{X}^k) - \left(\frac{1}{2t} - \frac{1}{2L_f}\right)\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2\\ &\leq f_{\max}^k - \left(\frac{1}{2t} - \frac{1}{2L_f}\right)\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2\\ &\leq f_{\max}^k - \frac{\delta}{2}\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2,\end{aligned}$$

provided that $\delta \leq \frac{1}{t} - \frac{1}{L_f}$. By defining $T_\delta = \min(\frac{\underline{\sigma}}{4c}, \frac{L_f}{\delta L_f + 1})$, we conclude that

$$f(\mathbf{X}^{k+1}) \leq f_{\max}^{k-1} - \frac{\delta}{2}\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2 \text{ for all } t \in [0, T_\delta). \quad (33)$$

In other words, the sufficient decrease condition in line 29 holds for all $t \in [0, T_\delta)$.

Letting T_p^k be the appropriate T_p^k from either Case 1 or Case 2 and letting σ and $\alpha_{k,0}$ be taken from Algorithm 1, we conclude that within $\lceil \log_\sigma(\frac{\min\{T_p^k, T_\delta\}}{\alpha_{k,0}}) \rceil$ iterations, $\mathbf{X}^{k+1}$ will satisfy both of the stopping criteria of the line search. □

Lemma 5. *Let $\{\mathbf{X}^k\}$ denote the sequence of iterations generated by Algorithm 1 for (5). The sequence $\{f(\mathbf{X}^k)\}$ admits a limit point f^*. Moreover, if the stepsize α_k is uniformly bounded away from zero, then it must hold $\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F \to 0$.*

Proof. If $\nabla f(\mathbf{X}^k) = \mathbf{0}$ for any k, then the lemma is immediate. So, suppose $\nabla f(\mathbf{X}^k) \neq \mathbf{0}$ for every k. Because of the stopping criteria for the line search in lines 16 and 29 of Algorithm 1, we have $f(\mathbf{X}^{k+1}) \leq f_{\max}^k$ for all k. Thus, $f_{\max}^{k+1} \leq f_{\max}^k$ for all k. The sequence $\{f_{\max}^k\}$ is monotone decreasing and bounded from below by $f(\mathbf{X}^*)$; hence, there exists a limit point $\{f_{\max}^k\} \to f^*$. Because f is convex and Lipschitz continuously differentiable on the set $\underline{\sigma}\mathbf{I}_n \leq \mathbf{X} \leq \overline{\sigma}\mathbf{I}_n\}$, we use the same induction argument as in Wright et al. 2009, lemma 4 to conclude that $\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F \to 0$. □

Theorem 4. *If $\mathbf{X}^*$ is an accumulation point of $\{\mathbf{X}^k\}$ generated by Algorithm 1, then $\mathbf{X}^*$ satisfies the necessary optimality conditions stated in Lemma 2. In other words, there exists $T > 0$ such that (19) holds, that is,*

$$\mathbf{X}^* = P_\Omega(\mathbf{X}^* - t\nabla f(\mathbf{X}^*)) \quad \text{for all} \quad t \in (0, T).$$

Proof. Once again, we will assume $\nabla f(\mathbf{X}^k) \neq \mathbf{0}$ for every k, because otherwise the theorem is immediate. Because $\mathbf{X}^*$ is an accumulation point of $\{\mathbf{X}^k\}$, there exists a subsequence of indices $\{k_i\}_{i=0,1,2,\ldots}$ such that $\lim_{i\to\infty} \mathbf{X}^{k_i} = \mathbf{X}^*$.

We now prove that $\{\alpha_{k_i}\}$ is uniformly bounded away from zero; in other words, $\alpha_{k_i} \geq \overline{\alpha} > 0$ for some $\overline{\alpha}$. First, suppose that there is no subsequence $|\{X_{m,n}^{k_i}|\}$ such that $\liminf_{i\to\infty} |X_{m,n}^{k_i}| = 0$ for some $(m, n) \in \cap_{k_i > \overline{k}} \mathcal{I}_{k_i}$ and $\overline{k} > 0$. Thus, for a sufficiently large $\overline{k}$, the sparsity patterns of $\mathbf{X}^{k_i}$ are the same with that of $\mathbf{X}^*$ for $k_i > \overline{k}$. We can claim that $\{|X_{i_x,j_x}^{k_i}|\}$ defined in the proof of Case 1 in Lemma 4 is uniformly bounded away from zero. Combining this fact with (32) and $\mathbf{x}^{k_i} \in \mathcal{L} \subset \{\mathbf{X} \in \mathbb{R}^{n\times n} : \underline{\sigma}\mathbf{I}_n \leq \mathbf{X} \leq \overline{\sigma}\mathbf{I}_n\}$, we have that there exists a constant $\overline{T}_p > 0$ such that $T_p^{k_i} = \min(T_1^{k_i}, T_2^{k_i}) \geq \overline{T}_p$ for all $k_i > \overline{k}$. Thus, the stepsize α_{k_i} taken at the end of the k_ith iteration is lower bounded:

$$\alpha_{k_i} \geq \alpha_{k_i,0}\sigma^{\left\lceil \log_\sigma\left(\frac{\min\{\overline{T}_p, T_\delta\}}{\alpha_{k_i,0}}\right)\right\rceil},$$

and so there exists $\overline{\alpha} > 0$ such that

$$\lim_{i\to\infty} \alpha_{k_i} \geq \overline{\alpha},$$

because we imposed bounds $\alpha_{\min} \leq \alpha_{k_i,0} \leq \alpha_{\max}$ as algorithmic parameters.

We consider the case when the set

$$\hat{\mathcal{I}}_\infty = \Big\{(m, n) : \liminf_{i\to\infty} |X_{m,n}^{k_i}| = 0, (m, n) \in \mathcal{I}_{k_i} \text{ for all } k_i \geq \overline{k} \text{ with some } \overline{k} > 0\Big\}$$

is nonempty. We assume for the sake of contradiction that $\alpha_{k_i} \to 0$. Define the swapping index sets through the projection step

$$\begin{aligned}\hat{\mathcal{I}}_k(\alpha) &= \left\{(i, j) \in \mathcal{I}_k : \left(P_\Omega(\mathbf{X}^k - \alpha\nabla f(\mathbf{X}^k))\right)_{(i,j)} = 0\right\}\\ \hat{\mathcal{A}}_k(\alpha) &= \left\{(i, j) \in \mathcal{A}_k : \left(P_\Omega(\mathbf{X}^k - \alpha\nabla f(\mathbf{X}^k))\right)_{(i,j)} \neq 0\right\}.\end{aligned}$$

By the line search procedure defined in line 26 and $\alpha_{k_i} \to 0$, we have that $\overline{\mathbf{X}}^{k_i+1} = P_\Omega(\mathbf{X}^{k_i} - \frac{\alpha_{k_i}}{\sigma}\nabla f(\mathbf{X}^{k_i}))$ violates at least one of conditions (C1) or (C2). Note that

f is Lipschitz continuously differentiable on $\mathcal{H}$ (defined in (31)) with a constant L_f. If $\frac{\alpha_{k_i}}{\sigma} < T_\delta = \min(\frac{\sigma}{4c}, \frac{L_f}{\delta L_f + 1})$, then (C1) is satisfied; hence, (C2) is violated by $\overline{\mathbf{X}}^{k_i+1}$. We can decompose

$$\overline{\mathbf{X}}^{k_i+1} = \mathbf{X}^{k_i} - \left[\mathbf{X}^{k_i}\right]_{\hat{\mathcal{I}}_{k_i}(\alpha_{k_i}/\delta)} - \frac{\alpha_{k_i}}{\delta}\left(\left[\nabla f(\mathbf{X}^{k_i})\right]_{\mathcal{I}_{k_i}\setminus\hat{\mathcal{I}}_{k_i}(\alpha_{k_i}/\delta)} + \left[\nabla f(\mathbf{X}^{k_i})\right]_{\hat{\mathcal{A}}_k(\alpha_{k_i}/\delta)}\right). \tag{34}$$

For a sufficiently large $\overline{k}$, it follows that $\hat{\mathcal{I}}_{k_i}(\alpha_{k_i}/\delta) \subseteq \hat{\mathcal{I}}_\infty$. Thus $[\mathbf{X}^{k_i}]_{\hat{\mathcal{I}}_{k_i}(\alpha_{k_i}/\delta)} \to \mathbf{0}$. Because $\{\nabla f(\mathbf{X}_k)\}$ is bounded, the third term in (34) vanishes if $i \to \infty$. Note that $\mathbf{X}^{k_i} \in \mathcal{H}$. Hence, $\overline{\mathbf{X}}^{k_i+1}$ is positive definite when k_i is sufficiently large, which is a contradiction. This implies that there exists $\overline{\alpha} > 0$ such that $\alpha_{k_i} \geq \overline{\alpha}$.

In the remainder of the proof, let T be such that $0 < T < \min(\overline{\alpha}, \overline{T}_p)$. We consider two possible cases for the subsequence $\{\mathbf{X}^{k_i+1}\}$ and analyze them separately: either infinitely many $\mathbf{X}^{k_i+1}$ are generated in Step 2, or else only finitely many $\mathbf{X}^{k_i+1}$ are generated in Step 2.

Case 1. Suppose that infinitely many iterations in $\{\mathbf{X}^{k_i+1}\}_{i=0,1,\ldots}$ are generated in Step 2 (the gradient projection step). Then, without loss of generality, we can assume that the entire subsequence $\{\mathbf{X}^{k_i+1}\}_{i=0,1,\ldots}$ is generated in Step 2. By triangle inequality, we have

$$\lim_{i\to\infty} \|\mathbf{X}^{k_i+1} - \mathbf{X}^*\|_F \leq \lim_{i\to\infty} \|\mathbf{X}^{k_i+1} - \mathbf{X}^{k_i}\|_F + \lim_{i\to\infty} \|\mathbf{X}^{k_i} - \mathbf{X}^*\|_F,$$

and so we conclude from Lemma 5 and the choice of subsequence $\{k_i\}$ that $\lim_{i\to\infty} \mathbf{X}^{k_i+1} = \mathbf{X}^*$.

For any $t > 0$, let $\mathbf{Y}^{k_i} \in \mathcal{S}_{t,\mathbf{I}_n}(\mathbf{X}^{k_i})$. It follows from the definition of $\mathcal{S}_{t,\mathbf{I}_n}$ that

$$\operatorname{tr}\left(\nabla f(\mathbf{X}^{k_i})^\mathsf{T}(\mathbf{X}^{k_i+1} - \mathbf{X}^{k_i})\right) + \frac{1}{2t}\|\mathbf{X}^{k_i+1} - \mathbf{X}^{k_i}\|_F^2 \geq \operatorname{tr}\left(\nabla f(\mathbf{X}^{k_i})^\mathsf{T}(\mathbf{Y}^{k_i} - \mathbf{X}^{k_i})\right) + \frac{1}{2t}\|\mathbf{Y}^{k_i} - \mathbf{X}^{k_i}\|_F^2. \tag{35}$$

Because $\mathbf{X}^{k_i+1} \in \mathcal{S}_{\alpha_{k_i},\mathbf{I}_n}(\mathbf{X}^{k_i})$ for all k_i, we also have from the definition of $\mathcal{S}_{\alpha_{k_i},\mathbf{I}_n}$ that

$$\operatorname{tr}(\nabla f(\mathbf{X}^{k_i})\operatorname{tr}(\mathbf{Y}^{k_i} - \mathbf{X}^{k_i})) + \frac{1}{2\alpha_{k_i}}\|\mathbf{Y}^{k_i} - \mathbf{X}^{k_i}\|_F^2 \geq \operatorname{tr}(\nabla f(\mathbf{X}^{k_i})\operatorname{tr}(\mathbf{X}^{k_i+1} - \mathbf{X}^{k_i})) + \frac{1}{2\alpha_{k_i}}\|\mathbf{X}^{k_i+1} - \mathbf{X}^{k_i}\|_F^2. \tag{36}$$

Combining (35) and (36), we obtain for any $t > 0$

$$\operatorname{tr}\left(\nabla f(\mathbf{X}^{k_i})^\mathsf{T}(\mathbf{X}^{k_i+1} - \mathbf{X}^{k_i})\right) + \frac{1}{2t}\|\mathbf{X}^{k_i+1} - \mathbf{X}^{k_i}\|_F^2 \geq \operatorname{tr}\left(\nabla f(\mathbf{X}^{k_i})^\mathsf{T}(\mathbf{X}^{k_i+1} - \mathbf{X}^{k_i})\right) + \frac{1}{2\alpha_{k_i}}\|\mathbf{X}^{k_i+1} - \mathbf{X}^{k_i}\|_F^2 + \left(\frac{1}{2t} - \frac{1}{2\alpha_{k_i}}\right)\|\mathbf{Y}^{k_i} - \mathbf{X}^{k_i}\|_F^2. \tag{37}$$

If $t \in (0, T)$, then we conclude from (37) that $\lim_{i\to\infty} \|\mathbf{Y}^{k_i} - \mathbf{X}^{k_i}\|_F = 0$, because $\lim_{i\to\infty} \|\mathbf{X}^{k_i+1} - \mathbf{X}^{k_i}\|_F = 0$ by Lemma 5 and because $\frac{1}{2t} - \frac{1}{2\alpha_{k_i}} > 0$ by the choice of t. By triangle inequality,

$$\lim_{i\to\infty} \|\mathbf{Y}^{k_i} - \mathbf{X}^*\|_F \leq \lim_{i\to\infty} \|\mathbf{Y}^{k_i} - \mathbf{X}^{k_i}\|_F + \lim_{i\to\infty} \|\mathbf{X}^{k_i} - \mathbf{X}^*\|_F,$$

and, hence, $\lim_{i\to\infty} \mathbf{Y}^{k_i} \to \mathbf{X}^*$. Thus, by Proposition 1, we conclude that $\mathbf{X}^* \in P_\Omega(\mathbf{X}^* - t\nabla f(\mathbf{X}^*))$ for all $t \in (0, T)$.

Case 2. Now suppose that only finitely many elements of the subsequence $\{\mathbf{X}^{k_i+1}\}$ are generated in Step 2. Then, without loss of generality, we can assume that every iteration of the subsequence $\{\mathbf{X}^{k_i+1}\}_{i=0,1,\ldots}$ is generated in Step 1. Let $p, q \in \mathbb{Z}$ be as in the statement of Algorithm 1. By the pigeonhole principle, there exists an integer $m \in [0, p)$ such that $\operatorname{mod}(k_i + 1, q) = m$ for infinitely many k_i. Hence, one has $\operatorname{mod}(k_i + 1 + p - m, q) = p$ for infinitely many k_i; in other words, every iteration of the subsequence $\{\mathbf{X}^{k_i+1+p-m}\}$ is generated in Step 2. From the triangle inequality,

$$\lim_{i\to\infty} \|\mathbf{X}^{k_i+1+p-m} - \mathbf{X}^*\|_F \leq \lim_{i\to\infty} \sum_{j=0}^{p-m} \|\mathbf{X}^{k_i+j+1} - \mathbf{X}^{k_i+j}\|_F + \lim_{i\to\infty} \|\mathbf{X}^{k_i} - \mathbf{X}^*\|_F,$$

and so by Lemma 5 and the choice of the subsequence $\{\mathbf{X}^{k_i}\}$, $\mathbf{X}^{k_i+1+p-m} \to \mathbf{X}^*$.

From here, we can apply the same argument as in Case 1 to the subsequence $\{\mathbf{X}^{k_i+1+p-m}\}$ to obtain $\mathbf{X}^* \in P_\Omega(\mathbf{X}^* - t\nabla f(\mathbf{X}^*))$ for all $t \in (0, T)$.

We now need only to prove the uniqueness of $P_\Omega(\mathbf{X}^* - t\nabla f(\mathbf{X}^*))$. That is, we must show that $\mathbf{X}^* = P_\Omega(\mathbf{X}^* - t\nabla f(\mathbf{X}^*))$ as a set equality. We analyze two separate cases.

Case A ($\|\mathbf{X}^*\|_0 = \kappa$): Assume that $\mathbf{Z}^* \in P_\Omega(\mathbf{X}^* - t\nabla f(\mathbf{X}^*))$ for all $t \in (0, T)$. Toward a contradiction, suppose that $\mathbf{Z}^* \neq \mathbf{X}^*$ for some $t' \in (0, T)$. By the definition of P_Ω, we conclude that because $\mathbf{Z}^* \neq \mathbf{X}^*$, $|X^*_{i',j'}| \leq |Z^*_{l',m'}(t')|$, where we have defined

$$|X^*(i', j')| \triangleq \operatorname{argmin}\{|X^*_{u,v}| : (u, v) \in \mathcal{I}_\infty\}$$

and

$$|Z^*_{l',m'}(t)| \triangleq \operatorname{argmax}\left\{|(\mathbf{X}^* - t\nabla f(\mathbf{X}^*))_{u,v}| : (u, v) \in \mathcal{A}_\infty\right\},$$

where $\mathcal{I}_\infty = \{(i, j) : X^*_{i,j} \neq 0\}$ and $\mathcal{A}_\infty = \{(i, j) : X^*_{i,j} = 0\}$. This implies that $(\nabla f(\mathbf{X}^*))_{l',m'} = 0$, because otherwise there exists $\tilde{t} \in (0, T)$ such that $|(\mathbf{X}^* - \tilde{t}\nabla f(\mathbf{X}^*))_{l',m'}| > |X^*_{i',j'}|$, which would contradict $\mathbf{X}^* \in P_\Omega(\mathbf{X}^* - \tilde{t}\nabla f(\mathbf{X}^*))$, because $(i', j') \in \mathcal{I}_\infty$ and $(l', m') \in \mathcal{A}_\infty$. But, if $(\nabla f(\mathbf{X}^*))_{l',m'} = 0$, then $\mathbf{X}^* \notin P_\Omega(\mathbf{X}^* - t\nabla f(\mathbf{X}^*))$ for any $t \in (0, T)$, because $(\mathbf{X}^* - t\nabla f(\mathbf{X}^*))_{l',m'} = (X^*)_{l',m'}$ and $(l', m') \in \mathcal{A}_\infty$. This is a contradiction.

Case B ($\|\mathbf{X}^*\|_0 < \kappa$). If $(\nabla f(\mathbf{X}^*))_{i,j} \neq 0$ for some $(i, j) \in \mathcal{A}_\infty$, then there exists $\tilde{t} \in (0, T)$ such that

$|(\mathbf{X}^* - \tilde{t}\nabla f(\mathbf{X}^*))_{i,j}| \neq 0$; hence, $\mathbf{X}^* \notin P_\Omega(\mathbf{X}^* - \tilde{t}\nabla f(\mathbf{X}^*))$, which would be a contradiction. Thus, it must be the case that $[\nabla f(\mathbf{X}^*)]_{\mathcal{A}_\infty} = \mathbf{0}$. Consider the reduced projection problem over the inactive indices in $\mathcal{I}_\infty$,

$$\min_{\mathbf{Y}} \sum_{(i,j)\in\mathcal{I}_\infty} \left(Y_{i,j} - \left(\mathbf{X}^* - t\nabla f(\mathbf{X}^*)\right)_{i,j}\right)^2. \tag{38}$$

Because $X^*_{i,j}$ is a solution to (38), we conclude that $[\nabla f(\mathbf{X}^*)]_{\mathcal{I}_\infty} = \mathbf{0}$; thus $\nabla f(\mathbf{X}^*) = \mathbf{0}$, and hence $\mathbf{X}^* = P_\Omega(\mathbf{X}^* - t\nabla f(\mathbf{X}^*))$ for any $t \in (0, T)$. □

Our next result establishes that any accumulation point produced by Algorithm 1 is in fact a strict local minimizer of (5).

Theorem 5. *Any accumulation point $\mathbf{X}^*$ of the sequence $\{\mathbf{X}^k\}$ generated by Algorithm 1 is a strict local minimizer of (5).*

Proof. We need to prove that there exists $\epsilon > 0$ such that the following holds; if $\mathbf{Y}$ is such that $\mathbf{X}^* + \mathbf{Y}$ is feasible with respect to the constraints of (5) and $\|\mathbf{Y}\|_F \in (0, \epsilon)$, then $f(\mathbf{X}^* + \mathbf{Y}) > f(\mathbf{X}^*)$.

From Theorem 4, there exists $T > 0$ such that $\mathbf{X}^* = P_\Omega(\mathbf{X}^* - t\nabla f(\mathbf{X}^*))$ for all $t \in (0, T)$. Equivalently, for all such t, the matrix $\mathbf{0}$ is the unique global minimizer of

$$\arg\min_{\mathbf{Y}:\mathbf{X}^*+\mathbf{Y}\in\Omega} g_t(\mathbf{Y}) \triangleq f(\mathbf{X}^*) + \operatorname{tr}\left(\nabla f(\mathbf{X}^*)\mathbf{Y}\right) + \frac{1}{2t}\|\mathbf{Y}\|_F^2.$$

Moreover, f is strongly convex on $\mathcal{H}$, as defined in (31). Let ρ denote the strong convexity constant. Note that $\mathbf{X}^k \in \mathcal{L} \subset \{\underline{\sigma}\mathbf{I}_n \preceq \mathbf{X} \preceq \overline{\sigma}\mathbf{I}_n\}$, and so the limit $\mathbf{X}^* \in \{\underline{\sigma}\mathbf{I}_n \preceq \mathbf{X} \preceq \overline{\sigma}\mathbf{I}_n\}$. We proved in Lemma 4 that if $\|\mathbf{Y}\|_F \leq 0.5\underline{\sigma}$, then $\mathbf{X}^* + \mathbf{Y} \in \mathcal{H}$. For any $\|\mathbf{Y}\|_F \leq 0.5\underline{\sigma}$, we have

$$\begin{aligned} f(\mathbf{X}^* + \mathbf{Y}) &\geq f(\mathbf{X}^*) + \operatorname{tr}(\langle \nabla f(\mathbf{X}^*), \mathbf{Y}\rangle) + \frac{1}{2\rho}\|\mathbf{Y}\|_F^2 \\ &= f(\mathbf{X}^*) + \frac{1}{2\rho}\|\mathbf{Y} + \rho\nabla f(\mathbf{X}^*)\|_F^2 - \frac{\rho}{2}\|\nabla f(\mathbf{X}^*)\|_F^2, \end{aligned} \tag{39}$$

where the equality follows from completing the square. As in the proof of Theorem 4, we now split the analysis into two cases.

Case 1 ($\|\mathbf{X}^*\|_0 = \kappa$). Observe that we must have $Y_{i,j} = 0$ for every $(i,j) \in \mathcal{A}_\infty = \{(i,j) : X^*_{i,j} = 0\}$, in order for $\mathbf{X}^* + \mathbf{Y} \in \Omega$ to hold in an arbitrarily small neighborhood of $\mathbf{X}^*$. Also observe that $(\nabla f(\mathbf{X}^*))_{i,j} = 0$ for every $(i,j) \in \mathcal{I}_\infty = \{(i,j) : X^*_{i,j} \neq 0\}$; otherwise, $\mathbf{X}^* \notin P_\Omega(\mathbf{X}^* - t\nabla f(\mathbf{X}^*))$ for sufficiently small t, which would yield a contradiction. Taking

$$\epsilon < \min\left\{0.5\underline{\sigma}, \min\left\{|X^*_{i,j}| : (i,j) \in \mathcal{I}_\infty\right\}\right\}, \tag{40}$$

we have $\mathbf{X}^* + \mathbf{Y} > 0$ because $0.5\underline{\sigma} < \sigma_1(\mathbf{X}^*)$. From (39), we obtain

$$\begin{aligned} f(\mathbf{X}^* + \mathbf{Y}) &\geq f(\mathbf{X}^*) + \frac{1}{2\rho}\sum_{(i,j)\in\mathcal{I}_\infty}\left(Y_{i,j} + \rho\left(\nabla f(\mathbf{X}^*)\right)_{i,j}\right)^2 \\ &\quad + \frac{1}{2\rho}\sum_{(i,j)\in\mathcal{A}_\infty}\left(Y_{i,j} + \rho\left(\nabla f(\mathbf{X}^*)\right)_{i,j}\right)^2 \\ &\quad - \frac{\rho}{2}\sum_{(i,j)\in\mathcal{I}_\infty}\left(\left(\nabla f(\mathbf{X}^*)\right)_{i,j}\right)^2 \\ &\quad - \frac{\rho}{2}\sum_{(i,j)\in\mathcal{A}_\infty}\left(\left(\nabla f(\mathbf{X}^*)\right)_{i,j}\right)^2 \\ &= f(\mathbf{X}^*) + \frac{1}{2\rho}\sum_{(i,j)\in\mathcal{I}_\infty} Y_{i,j}^2 \\ &\quad + \frac{1}{2\rho}\sum_{(i,j)\in\mathcal{A}_\infty}\left(\rho\left(\nabla f(\mathbf{X}^*)\right)_{i,j}\right)^2 \\ &\quad - \frac{\rho}{2}\sum_{(i,j)\in\mathcal{A}_\infty}\left(\left(\nabla f(\mathbf{X}^*)\right)_{i,j}\right)^2 \\ &= f(\mathbf{X}^*) + \frac{1}{2\rho}\sum_{(i,j)\in\mathcal{I}_\infty} Y_{i,j}^2 > f(\mathbf{X}^*), \end{aligned}$$

where the last strict inequality comes from the fact that not all the entries of $\mathbf{Y}$ are zero.

Case 2 ($\|\mathbf{X}^*\|_0 < \kappa$). As in the proof of Case 1 in Theorem 4, we get $\nabla f(\mathbf{X}^*) = \mathbf{0}$. Hence, from (39), we get

$$f(\mathbf{X}^* + \mathbf{Y}) \geq f(\mathbf{X}^*) + \frac{1}{2\rho}\|\mathbf{Y}\|_F^2 > f(\mathbf{X}^*)$$

for any $\|\mathbf{Y}\|_F \leq \epsilon$ (ϵ defined in (40)), which completes the proof. □

We now prove that a slightly modified version of the homotopy method HANA converges to a critical point of the Problem (3). The key modification is that whereas Algorithm 2 is stated (and implemented) with a finite list of parameters $\{(\kappa_\ell, \lambda_\ell)\}_{\ell=1,\ldots,L}$, we analyze Algorithm 2 with an infinite sequence of parameters $\{(\kappa_\ell, \lambda_\ell)\}_{\ell=0,1,\ldots}$ satisfying $(\kappa_\ell, \lambda_\ell) \to (\kappa, 0)$.

Theorem 6. *Assume the level set for Problem (3)*

$$\mathcal{L}_0 = \left\{\mathbf{X} \in \Omega \cap \mathbb{S}^n_{++} : f_0(\mathbf{X}) \leq f_0(\mathbf{X}^{(0)})\right\}$$

is a compact set for some $\mathbf{X}^{(0)} \in \Omega \cap \mathbb{S}^n_{++}$, where f_0 is defined in (6). Then

i. *Problem (3) has at least one global minimizer.*

ii. *Consider an infinite sequence of tuples $\{(\kappa_\ell, \lambda_\ell)\}_{\ell=0,1,\ldots}$ satisfying $(\kappa_\ell, \lambda_\ell) \to (\kappa, 0)$. Then, any accumulation point $\overline{\mathbf{X}}$ of $\{\mathbf{X}^{(\ell)}\}$ generated by Algorithm 2 for fixed pairs $(\kappa_\ell, \lambda_\ell)$ is a stationary point for Problem (3) satisfying Lemma 3 for some $T > 0$.*

Proof.

i. Immediate from the Weierstrass theorem.

ii. Without loss of generality (because κ only takes integer values), we can assume that $\kappa_\ell = \kappa$ for all ℓ. We explicitly define the objective function of (5) as $f_\lambda(\mathbf{X}) = f_0(\mathbf{X}) + \frac{\lambda}{2}\|\mathbf{X}\|_F^2$, which is parameterized by λ. From the assumed compactness of $\mathcal{L}_0$, there exists $M > 0$ such that $\|\mathbf{X}\|_F \le M$ for all $\mathbf{X} \in \mathcal{L}_0$. Define

$$\mathcal{L}_\lambda = \{\mathbf{X} \in \Omega \cap \mathbb{S}^n_{++} : f_\lambda(\mathbf{X}) \le f_\lambda(\mathbf{X}^{(0)})\}.$$

We show that $\|\mathbf{X}\|_F \le M$ for all $\mathbf{X} \in \mathcal{L}_\lambda$ for every $\lambda > 0$. Suppose for contradiction that there exists some $\mathbf{X} \in \mathcal{L}_\lambda$ such that $\|\mathbf{X}\|_F > M$. Hence, it follows that $\mathbf{X} \notin \mathcal{L}_0$, which implies $f_0(\mathbf{X}) > f_0(\mathbf{X}^{(0)})$. We have

$$\begin{aligned} f_\lambda(\mathbf{X}^{(0)}) &= f_0(\mathbf{X}^{(0)}) + \frac{\lambda}{2}\|\mathbf{X}^{(0)}\|_F^2 \\ &\le f_0(\mathbf{X}^{(0)}) + \frac{\lambda}{2}M^2 < f_0(\mathbf{X}) + \frac{\lambda}{2}\|\mathbf{X}\|_F^2 = f_\lambda(\mathbf{X}), \end{aligned}$$

which contradicts the assumption that $\mathbf{X} \in \mathcal{L}_\lambda$.

Next we show that there exist $0 < \underline{\sigma} < \overline{\sigma} < \infty$ such that $\underline{\sigma}\mathbf{I}_n \prec \mathbf{X} \prec \overline{\sigma}\mathbf{I}_n$ for any $\mathbf{X} \in \mathcal{L}_\lambda$ and $0 \le \lambda \le \lambda_0$. Because $\mathbf{X} \in \mathcal{L}_\lambda$, we have

$$M^2 \ge \|\mathbf{X}\|_F^2 \ge \frac{\sigma_{\max}^2(\mathbf{X})}{n}.$$

Hence, $\sigma_{\max}(\mathbf{X}) \le M\sqrt{n} = \overline{\sigma}$. To derive a lower bound, we have

$$\begin{aligned} f_0(\mathbf{X}^{(0)}) + \frac{\lambda_0}{2}\|\mathbf{X}^{(0)}\|_F^2 &\ge f_0(\mathbf{X}^{(0)}) + \frac{\lambda}{2}\|\mathbf{X}^{(0)}\|_F^2 \\ &\ge f_0(\mathbf{X}) + \frac{\lambda}{2}\|\mathbf{X}\|_F^2 \ge -\log\det(\mathbf{X}) \\ &\ge -\log(\sigma_{\min}(\mathbf{X})) - (n-1)\log(\overline{\sigma}). \end{aligned}$$

We get that $\sigma_{\min}(\mathbf{X}) \ge \underline{\sigma} = \exp\left(-f_0(\mathbf{X}^{(0)}) - \frac{\lambda_0}{2}\|\mathbf{X}^{(0)}\|_F^2\right)\overline{\sigma}^{1-n}$.

Because $\nabla^2 f_\lambda(\mathbf{X}) = \mathbf{X}^{-1} \otimes \mathbf{X}^{-1} + \lambda\mathbf{I}_{n^2}$, we have that $\nabla^2 f_\lambda(\mathbf{X}) \preceq ((0.5\underline{\sigma})^{-2} + \lambda_0)\mathbf{I}_{n^2}$ for every $\mathbf{X} \in \mathcal{H}$, where $\mathcal{H} = \{\mathbf{X} \in \mathbb{R}^{n\times n} : 0.5\underline{\sigma}\mathbf{I}_n \preceq \mathbf{X} \preceq (\overline{\sigma} + 0.5\underline{\sigma})\mathbf{I}_n\}$. Hence, a Lipschitz constant for f_λ over $\mathcal{H}$ is $L_{f_\lambda} = ((0.5\underline{\sigma})^{-2} + \lambda_0)^{-1}$.

We show that there exists $T > 0$ such that the optimality condition (19) holds for $\mathbf{X}^{(\ell)}$. Let $\mathbf{X}^{(\ell)k}$ denote the matrix generated by Algorithm 1 at the k-th iteration for the corresponding $(\kappa_\ell, \lambda_\ell)$. Define

$$\begin{aligned} \hat{\mathcal{J}}^\infty &= \Big\{(i,j) : \liminf_{i\to\infty} |X_{i,j}^{(\ell)}| = 0, (i,j) \in \mathcal{J}^{(\ell)} \text{ for all} \\ &\qquad \ell \ge \overline{\ell} \text{ with some } \overline{\ell} > 0\Big\} \\ \mathcal{J}^\infty &= \{(i,j) : \overline{X}_{ij} \ne 0\}, \text{ where } \mathcal{J}^{(\ell)} = \left\{(i,j) : X_{i,j}^{(\ell)} \ne 0\right\}. \end{aligned}$$

We consider two cases:

Case 1 ($\hat{\mathcal{J}}^\infty = \emptyset$). Recall that $\{|X^{(\ell)k}_{i_x j_x}|\}$, as defined in Lemma 4, is uniformly bounded away from zero. Because $\sigma_{\min}(\mathbf{X}^{(\ell)k}) \ge \underline{\sigma}$, because L_{f_λ} is a constant with respect to λ, and applying the definitions of T_p^k and T_δ from the proof of Lemma 4, the existence of such $T > 0$ is guaranteed.

Case 2 ($\hat{\mathcal{J}}^\infty \ne \emptyset$). We can choose ℓ sufficiently large so that every entry in $[\mathbf{X}^{(\ell)}]_{\hat{\mathcal{J}}^\infty}$ is significantly dominated by any element of $[\mathbf{X}^{(\ell)}]_{\mathcal{J}^\infty}$. Precisely, we can select ℓ such that $\|[\mathbf{X}^{(\ell)} - \overline{\mathbf{X}}]_{\mathcal{J}^\infty}\|_F \le \epsilon$ and $\|[\mathbf{X}^{(\ell)}]_{\hat{\mathcal{J}}^\infty}\|_F \le \epsilon$ for a sufficiently small $\epsilon = 0.1\min\{\frac{\underline{\sigma}}{n}, \min\{|\overline{\mathbf{X}}_{ij}| : (i,j) \in \mathcal{J}^\infty\}\}$. The gradient projection update step of Algorithm 1 has the following form, which is a similar decomposition to that in (34):

$$\begin{aligned} \mathbf{X}^{(\ell)k+1} = \mathbf{X}^{(\ell)k} &- \left[\mathbf{X}^{(\ell)k}\right]_{\hat{\mathcal{J}}_k(\alpha_k^\ell)} - \alpha_k^\ell\Big(\left[\nabla f_{\lambda_\ell}\left(\mathbf{X}^{(\ell)k}\right)\right]_{\mathcal{J}_k\setminus\hat{\mathcal{J}}_k(\alpha_k^\ell)} \\ &+ \left[\nabla f_{\lambda_\ell}\left(\mathbf{X}^{(\ell)k}\right)\right]_{\hat{\mathcal{A}}_k(\alpha_k^\ell)}\Big), \end{aligned} \tag{41}$$

where $\mathcal{J}_k, \hat{\mathcal{J}}_k(\alpha_k^\ell)$ and $\hat{\mathcal{A}}_k(\alpha_k^\ell)$ have been defined in the proof of Lemma 5 and α_k^ℓ is the stepsize. When k is sufficiently large, it follows that $\hat{\mathcal{J}}_k(\alpha_k^\ell) \subseteq \hat{\mathcal{J}}^\infty$. Hence, every element of $[\mathbf{X}^{(\ell)k}]_{\mathcal{J}_k\setminus\hat{\mathcal{J}}_k(\alpha_k^\ell)}$ dominates any element of $[\mathbf{X}^{(\ell)k}]_{\hat{\mathcal{J}}_k(\alpha_k^\ell)}$. Because $\|\mathbf{X}^{(\ell)k}\|_F \le M$, we have that $\nabla f_{\lambda_\ell}(\mathbf{X}^{(\ell)k})$ is bounded. Thus, $\{\alpha_k^\ell\}$ is uniformly bounded away from zero, which implies $T > 0$.

We have that for any $0 < t < T$,

$$\mathbf{X}^{(\ell)} = P_{\Omega_\ell}(\mathbf{X}^{(\ell)} - t\nabla f_{\lambda_\ell}(\mathbf{X}^{(\ell)}). \tag{42}$$

Because $\Omega_\ell = \Omega$ for any sufficiently large ℓ and $\lim_{\lambda_\ell\to 0} f_{\lambda_\ell}(\mathbf{X}) = f_0(\mathbf{X})$, taking the limit as ℓ tends to ∞ yields

$$\overline{\mathbf{X}} = P_\Omega(\overline{\mathbf{X}} - t\nabla f_0(\overline{\mathbf{X}})),$$

which completes the proof. □

5. Numerical Experiments

In this section, we present experimental results to evaluate the performance of our method, HANA, for solving the ℓ_0-constrained model (3) on synthetic and real-world data sets. We compare it with the penalty decomposition method (Lu and Zhang 2013) for solving (3) and the proximal point algorithm (PPA) (Wang et al. 2010) for solving the convex ℓ_1 model (2). The MATLAB code for PD was downloaded from http://people.math.sfu.ca/~zhaosong/ (used a version released in August 2010), whereas we used the code for PPA available at https://www.polyu.edu.hk/ama/profile/dfsun/index.html#Codes released on April 5, 2010. The MATLAB codes for PD and PPA were run with their default parameters. The data sets can be accessed at https://doi.org/10.1287/ijoc.2020.0991.

Our method HANA is implemented in MATLAB. We chose the following parameter values for ANA:

$$\sigma = 0.5, \delta = 10^{-6}, \epsilon = 10^{-2}, \alpha_{\min}^{Newton} = 10^{-5},$$
$$\alpha_{\min} = 10^{-15}, \alpha_{\max} = 10^{15}, M = 2, p = 1, q = 5.$$

In Algorithm 2, we set $L = 10, \kappa_0 = 2\kappa, \lambda_L = 10^{-5}$ and $\lambda_0 = 10^{-1}$. The values κ_ℓ were uniformly spaced on the interval $[\kappa, \kappa_0]$, and λ_ℓ is an element of a vector of K logarithmically spaced points in the interval $[\lambda_L, \lambda_0]$. Subproblem (25) was solved inexactly by the conjugate gradient algorithm (Olsen et al. 2012) and it was terminated as soon as $\|\nabla q(\mathbf{D}^k)\|_\infty \le 0.05$. The stopping criterion for ANA at the final step of HANA ($\ell = L$) was chosen as

$$|f(\mathbf{X}^k) - f(\mathbf{X}^{k-1})| \le 10^{-5} \max\{1, |f(\mathbf{X}^{k-1})|\}$$

and for the previous steps ($\ell = 0, \ldots, L-1$) is

$$|f(\mathbf{X}^k) - f(\mathbf{X}^{k-1})| \le 10^{-2} \max\{1, |f(\mathbf{X}^{k-1})|\}.$$

We used an initial point $\mathbf{X}^{(0)} = (\mathrm{diag}(\mathbf{S}) + .10^{-2}\mathbf{I}_n))^{-1}$ as suggested in Olsen et al. (2012). All experiments were carried out on a ThinkPad W540 laptop using MATLAB 8.3 with a 64-bit Windows 7 with Intel i7-4800MQ 2.7 GHz CPU and 16 GB of RAM. Because Problem (3) is nonconvex, we are interested not only in the running times but also in the the quality of solutions returned by these methods. We are particularly interested in comparing the solutions returned by nonconvex solvers applied to (3) with some baseline solution returned from solving the convex model (2) to global optimality. We assume throughout these experiments that none of the independence structure is known; in other words, we choose $\mathfrak{I} = \emptyset$ in (3).

5.1. Synthetic Data Sets

We created synthetic test data sets by generating a ground truth sparse inverse covariance matrix $\hat{\mathbf{S}}^{-1}$ in a number of different scenarios. We consider four types of graph structures, also considered in d'Aspremont et al. (2008), Li and Toh (2010), Lu and Zhang (2013), and Hsieh et al. (2014), which we briefly describe here:

- Chain graph: $\hat{S}^{-1}_{i,i} = 1, \hat{S}^{-1}_{i,i-1} = \hat{S}^{-1}_{i-1,i} = 1.25$ and zero otherwise;
- AR2: $\hat{S}^{-1}_{i,i} = 1, \hat{S}^{-1}_{i,i-1} = \hat{S}^{-1}_{i-1,i} = 0.5, \hat{S}^{-1}_{i,i-2} = \hat{S}^{-1}_{i-2,i} = 0.25$ and zero otherwise;
- AR3: $\hat{S}^{-1}_{i,i} = 1, \hat{S}^{-1}_{i,i-1} = \hat{S}^{-1}_{i-1,i} = 0.4, \hat{S}^{-1}_{i,i-2} = \hat{S}^{-1}_{i-2,i} = \hat{S}^{-1}_{i,i-3} = \hat{S}^{-1}_{i-3,i} = 0.2$ and zero otherwise;
- Random graph: The true random sparse inverse covariance matrix is generated in a similar manner as that described in d'Aspremont (2008), and Li and Toh (2010). The sparsity density was controlled so that the matrix $\hat{\mathbf{S}}^{-1}$ has approximately $15n$ nonzero elements.

We generated $2n$ i.i.d. random samples from the n-dimensional Gaussian distribution $\mathcal{N}(\mathbf{0}, \hat{\mathbf{S}})$ and computed the sample covariance matrix $\mathbf{S}$. When the ground truth inverse covariance matrix $\hat{\mathbf{S}}^{-1}$ is known, we can utilize two commonly used loss metrics to evaluate the quality of recovery achieved by the estimated matrix $\mathbf{X}$: the normalized entropy loss (lossE) and quadratic loss (lossQ) (Li and Toh 2010, Lu and Zhang 2013). These metrics are defined, respectively, as

$$\texttt{lossE} = \frac{1}{n}\left(\mathrm{tr}(\hat{\mathbf{S}}\mathbf{X}) - \log\det\mathbf{X} - n\right),$$
$$\texttt{lossQ} = \frac{1}{n}\|\hat{\mathbf{S}}\mathbf{X} - \mathbf{I}_n\|_F.$$

We also tested the accuracy of the sparsity pattern recovered in $\mathbf{X}$ by these optimization methods by comparing with the sparsity pattern of the ground truth $\hat{\mathbf{S}}$. This was done via the standard statistical measures of true negative rate (TNR) (specificity) and true positive rate (TPR) (sensitivity) (see, e.g., Li and Toh 2010),

$$\texttt{TNR} = \frac{\texttt{TN}}{\texttt{TN} + \texttt{FP}} = \frac{|\{(i,j) \ : \ X_{i,j} = 0, \hat{S}_{i,j} = 0\}|}{|\{(i,j) \ : \ \hat{S}_{i,j} = 0\}|},$$
$$\texttt{TPR} = \frac{\texttt{TP}}{\texttt{TP} + \texttt{FN}} = \frac{|\{(i,j) \ : \ X_{i,j} \neq 0, \hat{S}_{i,j} \neq 0\}|}{|\{(i,j) \ : \ \hat{S}_{i,j} \neq 0\}|},$$

where TP, TN, TN and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively. In general, a method is sought that returns both TNR and TPR as close to one as possible; and the smaller for lossE and lossQ the better. We also used the log-likelihood performance metric in the comparison, which is the negative of the objective function of (3) (denoted by "likelihood"):

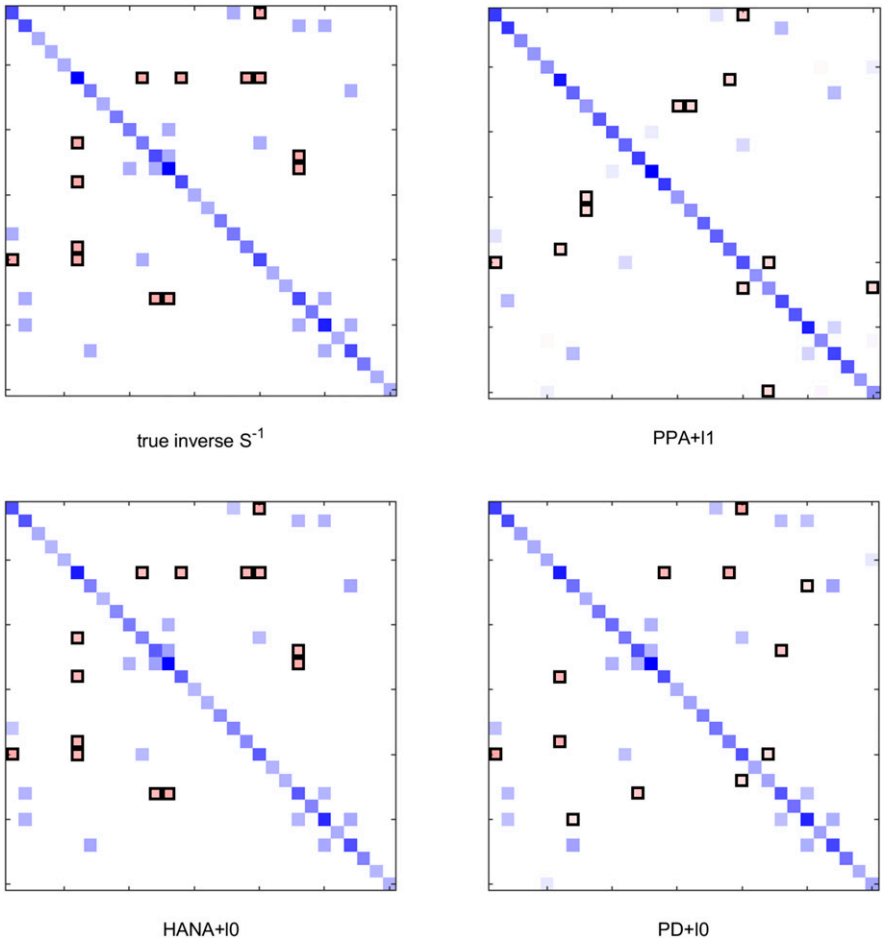
$$\texttt{likelihood} = -f_0(\mathbf{X}) = \log\det(\mathbf{X}) - \mathrm{tr}(\mathbf{S}\mathbf{X}).$$

When training the ℓ_1 model (2), we selected the penalty parameter λ so that the optimal solution had roughly the correct number of nonzero elements κ. A wide range of problem sizes n from $n = 500$ up to $n = 2{,}000$ was used, and the results are shown in Table 1. We observe that our method HANA is always faster than either the PD method or the PPA method. We note that these test instances have very sparse solutions (i.e., $\kappa \ll n^2$), and we surmise that our approximate Hessian-based method is handily exploiting this sparsity. Applied to the same nonconvex problem (3), HANA consistently identified a better solution in terms of the value of log-likelihood than the solution returned by PD; that is, HANA generated solutions with higher values for likelihood. The log-likelihoods attained by the ℓ_0 model (3) generated by both our method HANA and PD are significantly larger than those attained by

Table 1. The CPU Times in Seconds, Log-Likelihood, Losses, and True Negative/Positive Rates for HANA and PD Solving (3) and PPA Solving (2) for Synthetic Data Sets

a			HANA+ℓ_0a						PD+ℓ_0a						PPA+ℓ_1					
Data set	n	κ	Time	Likelihood	lossE	lossQ	TNR	TPR	Time	Likelihood	lossE	lossQ	TNR	TPR	Time	Likelihood	lossE	lossQ	TNR	TPR
Rand	500	8,170	3.22	103.67	0.0653	0.0196	0.9937	0.8132	4.29	91.00	0.0915	0.0213	0.9905	0.7200	7.58	23.62	0.1922	0.0235	0.9893	0.6842
Rand	1,000	16,848	15.55	212.96	0.0260	0.0086	0.9985	0.9149	25.04	184.56	0.0575	0.0116	0.9968	0.8137	29.13	59.28	0.1693	0.0156	0.9961	0.7706
Rand	1,500	25,324	46.79	322.57	0.0243	0.0068	0.9992	0.9269	65.51	275.75	0.0605	0.0098	0.9979	0.8180	109.29	79.54	0.1765	0.0130	0.9972	0.7600
Rand	2,000	34,160	108.51	421.76	0.0267	0.0061	0.9993	0.9180	148.10	360.19	0.0618	0.0084	0.9984	0.8190	231.07	88.39	0.1819	0.0113	0.9978	0.7522
AR2	500	2,494	3.47	−659.99	0.0032	0.0042	1	1	7.24	−660.24	0.0044	0.0049	1	0.9968	15.08	−758.83	0.1999	0.0215	0.9970	0.7113
AR2	1,000	4,994	16.72	−1,322.12	0.0015	0.0020	1	1	29.85	−1,322.32	0.0018	0.0022	1	1	68.09	−1,527.40	0.2053	0.0154	0.9983	0.6700
AR2	1,500	7,494	54.28	−1,983.06	0.0015	0.0016	1	1	122.07	−1,983.39	0.0018	0.0019	1	1	147.59	−2,290.30	0.2049	0.0125	0.9989	0.6685
AR2	2,000	9,994	112.52	−2,644.93	0.0015	0.0014	1	1	341.16	−2,645.51	0.0016	0.0020	1	1	455.39	−3,054.36	0.2049	0.0109	0.9992	0.6692
AR3	500	2,494	4.10	−604.31	0.0055	0.0052	0.9999	0.9953	7.06	−609.25	0.0194	0.0097	0.9988	0.9186	13.01	−642.18	0.0739	0.0141	0.9960	0.7133
AR3	1,000	4,994	19.99	−1,215.89	0.0020	0.0022	1	1	29.63	−1,216.71	0.0040	0.0031	0.9999	0.9906	61.48	−1,276.23	0.0610	0.0091	0.9980	0.7149
AR3	1,500	7,494	56.13	−1,823.73	0.0020	0.0018	1	1	120.38	−1,825.98	0.0054	0.0030	0.9999	0.9830	216.92	−1,915.82	0.0621	0.0075	0.9987	0.7145
AR3	2,000	9,994	123.01	−2,430.05	0.0020	0.0014	1	1	291.95	−2,437.18	0.0069	0.0029	0.9999	0.9747	433.73	−2,555.02	0.0622	0.0065	0.9990	0.7145
Chain	500	1,498	2.96	−498.11	0.0022	0.0034	1	1	4.56	−498.20	0.0026	0.0037	1	1	6.74	−604.53	0.2110	0.0232	1	0.9987
Chain	1,000	2,998	15.96	−997.67	0.0011	0.0017	1	1	24.10	−997.87	0.0013	0.0018	1	1	36.55	−1,203.46	0.2116	0.0164	1	1
Chain	1,500	4,498	53.03	−1,496.36	0.0010	0.0014	1	1	86.70	−1,496.70	0.0014	0.0015	1	1	102.13	−1,814.42	0.2117	0.0134	1	1
Chain	2,000	5,998	115.04	−1,995.96	0.0011	0.0012	1	1	236.03	−1,996.43	0.0013	0.0014	1	1	212.71	−2,419.85	0.2117	0.0116	1	1

Figure 1. (Color online) Sparsity Pattern Comparison for a Synthetic Data ($n = 30$)



solving the convex ℓ_1 model (2) when tuned to return a solution with similar κ nonzeros. This relative superiority of the ℓ_0 model over the ℓ_1 model in terms of log-likelihood also holds for the loss measures (lossE and lossQ) and statistical rates (TNR and TPR). The solutions returned by HANA applied to the ℓ_0-constrained model were consistently best in the metrics pertaining to the accuracy of recovery, whereas the solutions to the ℓ_1-regularized model were consistently the worst. Because of the better log-likelihood values of solutions returned by HANA than those of PD applied to (3), the solution returned by HANA has fewer misplaced nonzero elements in the sparsity pattern than do those solutions returned by PD; that is, the TNR and TPR values for solutions returned by HANA are either equal to, or very close to, one in many cases.

We provide a small-sized test example ($n = 30$) of a random graph to pictorially illustrate sparsity pattern recovery of the various methods, shown in Figure 1. The solution returned by HANA applied to the ℓ_0-model perfectly recovers the sparsity pattern of the ground truth inverse covariance matrix, whereas the solution of PD commits several errors. As shown in Table 2, PD became trapped in a local, but not a global, minimum; and so the log-likelihood of its solution is smaller than that of the solution returned by HANA. We observe that the solution returned by PPA applied to the ℓ_1- model misplaces many nonzero entries in this example.

In Table 3, we report the running times and the final objective function values when ANA and PD solve (5) for $\lambda = 0.01$. We can see that ANA solves these instances much faster than PD, and it achieves a better objective function value in five out of eight test instances. When embedding ANA into the homotopy framework, HANA outperforms PD for all eight cases.

Table 2. Results for a Synthetic Data ($n = 30$): Log-Likelihood, Losses, and True Negative/False Rates

Method	n	κ	Likelihood	lossE	lossQ	TNR	TPR
HANA+ℓ_0	30	62	−13.31	0.0032	0.0154	1	1
PD+ℓ_0	30	62	−13.55	0.0155	0.0328	0.9928	0.9032
PPA+ℓ_1	30	62	−15.07	0.0610	0.0531	0.9809	0.7419

Table 3. CPU Times in Seconds and Function Values for ANA and PD to Solve (5) with $\lambda = 0.01$ on Synthetic Data

			ANA+ℓ_0		PD+ℓ_0	
Data set	n	κ	Time	Function value	Time	Function value
Rand	500	8,170	1.34	−40.79	4.02	−46.03
Rand	1,000	16,848	7.44	−77.25	23.59	−52.17
Rand	1,500	25,324	23.80	−118.71	58.47	−81.66
Rand	2,000	34,160	56.83	−144.15	139.11	−138.08
AR3	500	2,494	2.10	612.07	6.53	620.25
AR3	1,000	4,994	9.14	1,223.16	26.04	1,219.94
AR3	1,500	7,494	31.52	1,839.85	117.08	1,839.85
AR3	2,000	9,994	76.45	2,447.05	274.55	2,509.73

5.2. Real-World Data Sets

In our second set of experiments, we consider five gene expression data sets that have been widely used in the literature (see, for example, Li and Toh et al. 2010, Lu and Zhang et al. 2013, Hsieh et al. 2014). Data preprocessing was performed via the same technique used in (Li and Toh 2010). We first solved the ℓ_1-model (2) with different regularization parameters $\lambda = 0.9, 0.7, 0.5, 0.3, 0.1, 0.05$ and 0.01. For each λ, we chose a corresponding value of κ for Problem (3) as the number of

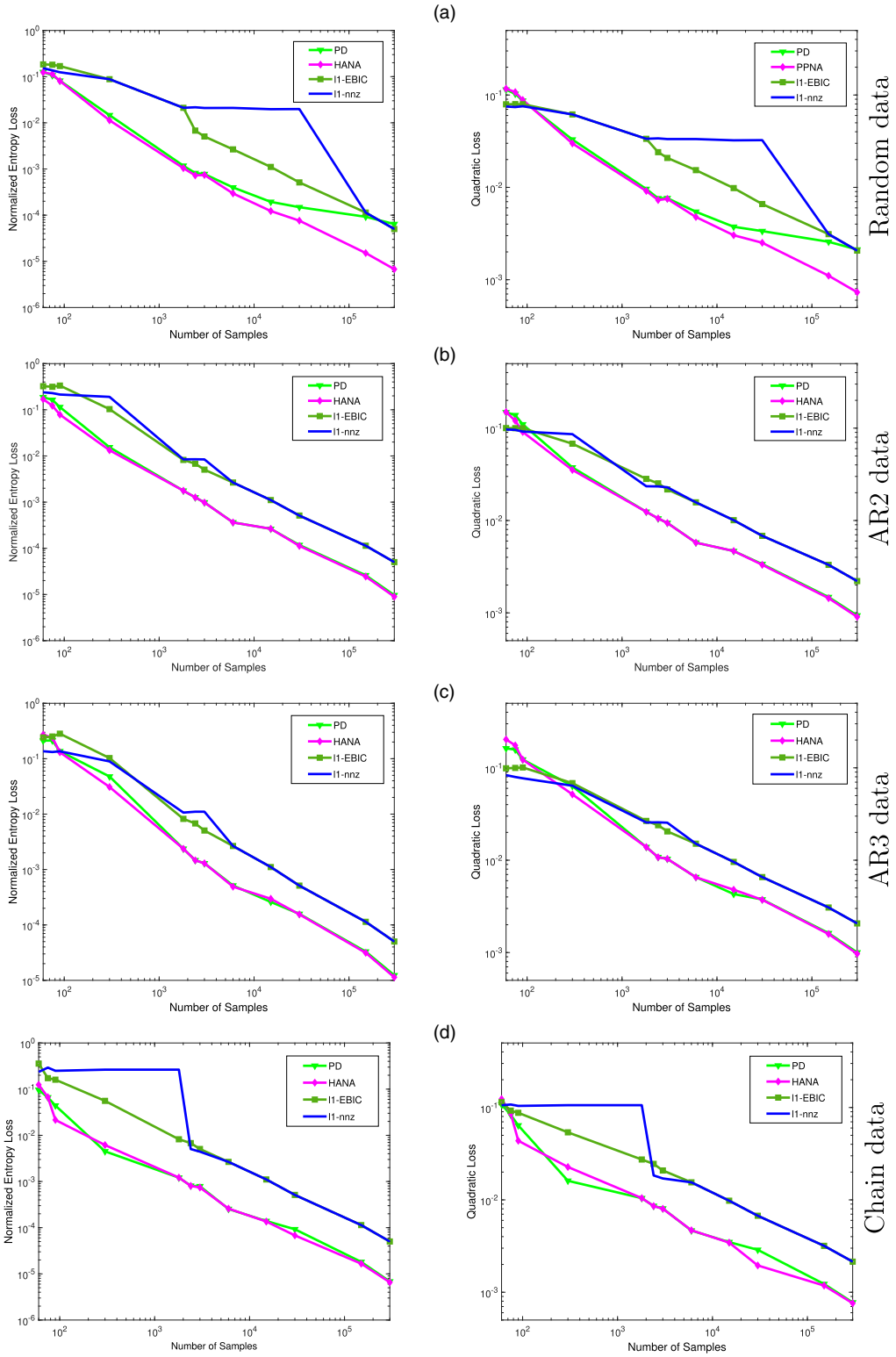
Table 4. Performance Comparison for Real Data Sets

				HANA+ℓ_0		PD+ℓ_0		PPA+ℓ_1	
Data set	n	λ	κ	Time	Likelihood	Time	Likelihood	Time	Likelihood
Lymph	587	0.9	713	8.62	−496.21	35.76	−526.59	11.50	−684.59
Lymph	587	0.7	1,443	11.48	−347.70	39.97	−422.48	13.68	−642.05
Lymph	587	0.5	4,607	18.61	−134.18	87.75	−213.51	26.96	−561.38
Lymph	587	0.3	14,873	19.30	109.81	105.50	79.23	31.49	−365.31
Lymph	587	0.1	37,599	25.26	524.43	27.17	405.22	32.31	−47.03
Lymph	587	0.05	67,503	35.42	854.63	18.16	714.37	37.55	174.86
Lymph	587	0.01	149,791	43.22	1,046.40	22.62	1,042.16	40.92	790.12
Estrogen	692	0.9	864	9.08	−528.68	65.49	−569.06	26.91	−806.84
Estrogen	692	0.7	5,494	45.06	196.69	165.34	146.89	25.64	−699.11
Estrogen	692	0.5	10,584	40.55	321.75	275.18	306.52	44.97	−516.28
Estrogen	692	0.3	16,882	43.18	449.28	140.64	425.00	58.98	−273.64
Estrogen	692	0.1	34,662	33.71	736.79	51.54	630.78	58.47	138.36
Estrogen	692	0.05	61,704	44.65	1,021.74	29.13	876.96	51.87	384.58
Estrogen	692	0.01	164,706	72.20	1,326.27	28.52	1,283.03	55.76	1,072.89
Arabidopsis	834	0.9	1,062	32.91	−535.22	135.72	−650.97	35.41	−970.57
Arabidopsis	834	0.7	6,610	97.10	278.42	597.96	−203.63	46.38	−856.47
Arabidopsis	834	0.5	20,540	88.74	680.79	481.68	531.72	62.17	−581.65
Arabidopsis	834	0.3	28,160	71.12	878.04	322.66	785.82	77.84	−275.11
Arabidopsis	834	0.1	51,754	90.79	1,382.25	68.73	1,150.99	88.11	309.68
Arabidopsis	834	0.05	84,346	92.34	1,675.81	48.20	1,487.47	82.23	692.52
Arabidopsis	834	0.01	193,082	115.35	1,899.66	42.97	1,837.38	90.13	1,718.41
Leukemia	1,225	0.9	1,279	32.76	−1,228.69	513.84	−1,239.25	106.24	−1,465.70
Leukemia	1,225	0.7	5,057	142.61	−249.94	1,681.06	−794.30	170.22	−1,367.22
Leukemia	1,225	0.5	34,657	290.26	1,237.78	1,889.67	1,009.97	267.64	−931.59
Leukemia	1,225	0.3	44,997	237.25	1,722.08	994.04	1,420.94	261.15	−489.58
Leukemia	1,225	0.1	102,361	203.16	2,927.66	140.27	2,531.33	199.59	505.02
Leukemia	1,225	0.05	167,881	463.31	3,315.66	117.22	2,962.68	193.58	1,308.39
Leukemia	1,225	0.01	272,451	368.67	3,320.26	112.67	3,172.67	232.40	3,229.76
Hereditary	1,869	0.9	2,017	110.91	−1,728.83	3,024.64	−1,794.24	402.24	−2,182.29
Hereditary	1,869	0.7	29,155	1,525.97	2,987.23	3,579.69	3,144.48	544.17	−1,771.88
Hereditary	1,869	0.5	69,159	714.97	5,236.13	590.04	4,251.42	618.43	−833.19
Hereditary	1,869	0.3	81,627	937.22	5,463.42	503.87	4,476.27	521.40	77.13
Hereditary	1,869	0.1	118,299	587.98	5,300.70	428.65	4,791.94	551.62	2,153.89
Hereditary	1,869	0.05	137,199	1,547.33	5,853.03	390.90	4,895.05	608.69	3,439.49
Hereditary	1,869	0.01	753,579	2,263.72	6,924.65	325.42	5,287.78	812.30	6,414.77

nonzero elements in the solution of (2) returned by PPA. The log-likelihood values and running times in seconds are summarized in Table 4.

When the density of solutions is low, which is especially the case when $\lambda \geq 0.3$, HANA is substantially faster than PD in terms of wall-clock time, except in the

Figure 2. (Color online) Number of Samples Versus Losses for a Synthetic Data ($n = 30$)



case of the data set Hereditary with $\lambda = 0.5$ and $\lambda = 0.3$. HANA is comparable to PPA in most cases in terms of wall-clock time but seems to perform better in accuracy on problems with very sparse solutions; only on problems with dense solutions does HANA's advantage disappear. This is to be expected, however, since we require only a small working set for computing Newton directions in our method. We observe that PD yields much better performance when the optimal solutions are fairly dense (e.g., when $\lambda \leq 0.1$). The performance of PPA applied to (2) seems to scale well with the density of the returned solution.

We observe that HANA performs best in terms of log-likelihood values. For a fixed number of nonzero elements, the log-likelihood values attained by HANA are the largest in all cases except for Hereditary with $\lambda = 0.7$, while solutions to the ℓ_1-model often provide the worst values of log-likelihood. In practice, we are interested mostly in cases where the solution is expected to be sparse, and so it encouraging that our proposed method is advantageous in terms of both wall-clock time and solution quality in the sparser settings tested. Because of the presence of the ℓ_0-constraint, it is widely perceived that solving the ℓ_0-model may be prohibitively expensive and a local search method could perform poorly. As shown in Tables 1 and 4, however, solutions to the ℓ_0-model returned by HANA consistently attain significantly higher log-likelihoods than do solutions to the ℓ_1 model, without incurring significant additional computation cost.

5.3. Empirical Convergence Rate for Consistency

In this subsection, we study the empirical convergence rate of consistency of the ℓ_0 model (3) and further demonstrate the relative benefit of solving model (3) as opposed to the ℓ_1 model (2). We consider the recovery error as a function of the number of sample points. It is known (see, e.g., Rothman et al. 2008) that solutions to (2) converge to the ground truth inverse covariance matrix at a rate $O_P(\sqrt{\frac{\hat{\kappa}\log n}{m}})$, where $\hat{\kappa}$ is the actual number of nonzeros and m is the sample size. We are not aware of any such theoretical consistency results for the ℓ_0 model.

We considered the same technique for tuning the value of the regularization parameter λ in the ℓ_1 model (2) based on values of κ as described in the Subsection 5.1. We denote this tuning method by "ℓ_1**-nnz**." We also tested another method for tuning the value of λ, wherein we chose a value of λ that approximately minimizes the extended Bayesian information criterion (EBIC) (Foygel and Drton 2010),

$$\texttt{EBIC}(\lambda) = \operatorname{tr}(\hat{\mathbf{S}}\mathbf{X}_\lambda) - \log\det \mathbf{X}_\lambda + \bar{\kappa}\frac{\log m}{m} + 2\bar{\kappa}\frac{\log n}{m}, \tag{43}$$

where $\mathbf{X}_\lambda$ is the estimate of the inverse covariance for a given λ and $\bar{\kappa}$ is the number of nonzero entries in the lower triangular part of $\mathbf{X}$. Denote "ℓ_1-EBIC" by this tuning method.

We used the same methods for generating synthetic data of size $n = 30$ discussed in Subsection 5.1. Figure 2 plots the normalized entropy loss (lossE) and quadratic loss (lossQ) versus the number of samples m. The solutions to (3) obtained by HANA and PD exhibit a faster convergence rate than do the solutions to the ℓ_1 model. That is, for a fixed level of desired accuracy, solutions to the ℓ_0 model generally recover something closer to the ground truth with fewer samples. This is an important feature in various applications where the number of samples is limited. We also observe that HANA produces slightly more robust results than does PD in these tests of empirical convergence rates.

6. Conclusions

In this paper, we have proposed a homotopy approximate Newton algorithm, HANA, to solve the ℓ_0-constrained inverse covariance learning problem. The algorithm is based on solving a sequence of bounded, ℓ_2-regularized subproblems by ANA. The method occasionally performs a line search in the projected gradient direction in the full space. This feature enables us to prove a global convergence result about the method as well as provides a practical means of identifying a good subspace of free variables in which we compute an approximate Newton update. We demonstrate that the algorithm HANA accumulates at a stationary point of the ℓ_0-constrained problem. In a series of numerical experiments, HANA found generally better solutions to the nonconvex problem (3) than did the PD and was more efficient on very sparse instances. Solutions returned by HANA often yield significantly higher values of log-likelihood than solutions to the convex ℓ_1 model with tuned values of the regularization parameter.

Acknowledgments

The authors thank Jayant Kalagnanam, Katya Scheinberg, and Peder Olsen for their valuable comments and suggestions during the course of this work.

Endnote

[1] The notation $\mathbf{X}^{-1}$ here is not to be confused with the standard notation for matrix inverse, as it is used everywhere else in this paper.

References

Barzilai J, Borwein JM (1988) Two point step size gradient methods. *IMA J. Numerical Anal.* 8:141–148.

Bertsimas D, King A, Mazumder R (2016) Best subset selection via a modern optimization lens. *Ann. Statist.* 44(2):813–852.

Birgin EG, Martínez JM, Raydan M (2000) Nonmonotone spectral projected gradient methods for convex sets. *SIAM J. Optim.* 10(4):1196–1211.

d'Aspremont A, Banerjee O, Ghaoui LE (2008) First-order methods for sparse covariance selection. *SIAM J. Matrix Anal. Appl.* 30(1):56–66.

Foygel R, Drton M (2010) Extended Bayesian information criteria for Gaussian graphical models. Lafferty JD, Williams CKI, eds. *Proc. 23rd Internat. Conf. Neural Inform. Processing Systems*, vol. 1 (Curran Associates, Red Hook, NY), 604–612.

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

Grippo L, Lampariello F, Lucidi S (1986) A nonmonotone line search technique for Newton's method. *SIAM J. Numerical Anal.* 23(4):707–716.

Hager WW, Phan DT, Zhu J (2016) Projection algorithms for nonconvex minimization with application to sparse principal components analysis. *J. Global Optim.* 65(4):657–676.

Hsieh CJ, Sustik MA, Dhillon IS, Ravikumar P (2014) QUIC: Quadratic approximation for sparse inverse covariance estimation. *J. Machine Learn. Res.* 15(83):2911–2947.

Johnson K, Lin D, Stine RA, Foster DP (2015) A risk ratio comparison of ℓ_0 and ℓ_1 penalized regression. Working paper, The Wharton School, University of Pennsylvania, Philadelphia.

Li L, Toh KC (2010) An inexact interior point method for ℓ_1-regularized sparse covariance selection. *Math. Programming Comput.* 2(3):291–315.

Liu Z, Lin S, Deng N, McGovern DP, Piantadosi S (2016) Sparse inverse covariance estimation with ℓ_0 penalty for network construction with omics data. *J. Comput. Biol.* 23(3):192–202.

Lu Z (2015) Optimization over sparse symmetric sets via a nonmonotone projected gradient method. Preprint, submitted September 29, https://arxiv.org/abs/1509.08581.

Lu Z, Zhang Y (2013) Sparse approximation via penalty decomposition methods. *SIAM J. Optim.* 23(4):2448–2478.

Marjanovic G, Hero AO (2015) ℓ_0 sparse inverse covariance estimation. *IEEE Trans. Signal Processing* 63(12):3218–3231.

Marjanovic G, Ulfarsson M, Solo V (2016) Large-scale ℓ_0 sparse inverse covariance estimation. *IEEE Internat. Conf. Acoustics Speech Signal Processing (ICASSP)* (IEEE, Piscataway, NJ), 4767–4771.

Olsen PA, Oztoprak F, Nocedal J, Rennie SJ (2012) Newton-like methods for sparse inverse covariance estimation. Pereira F, Burges CJC, Bottou L, eds. *Proc. 25th Internat. Conf. Neural Inform. Processing Systems* (Curran Associates, Red Hook, NY), 755–763.

Rothman AJ, Bickel PJ, Levina E, Zhu J (2008) Sparse permutation invariant covariance estimation. *Electronic J. Statist.* 2: 494–515.

Scheinberg K, Ma S, Goldfarb D (2010) Sparse inverse covariance selection via alternating linearization methods. Lafferty JD, Williams CKI, eds. *Proc. 23rd Internat. Conf. Neural Inform. Processing Systems* (Curran Associates, Red Hook, NY), 2101–2109.

Treister E, Turek J, Yavneh I (2016) A multilevel framework for sparse optimization with application to inverse covariance estimation and logistic regression. *SIAM J. Sci. Comput.* 38(5):S566–S592.

Wang C, Sun D, Toh KC (2010) Solving log-determinant optimization problems by a Newton-CG primal proximal point algorithm. *SIAM J. Optim.* 20(6):2994–3013.

Wright SJ, Nowak RD, Figueiredo MAT (2009) Sparse reconstruction by separable approximation. *IEEE Trans. Signal Processing* 57(7):2479–2493.

Xu F, Lu Z, Xu Z (2016) An efficient optimization approach for a cardinality-constrained index tracking problem. *Optim. Methods Software* 31(2):258–271.

Yuan X, Li P, Zhang T (2014) Gradient hard thresholding pursuit for sparsity-constrained optimization. *Proc. 31st Internat. Conf. Machine Learn.* (PMLR, *Beijing, China*), 127–135.

Yuan XT, Li P, Zhang T (2018) Gradient hard thresholding pursuit. *J. Machine Learn. Res.* 18(166):1–43.

Zhang RY, Fattahi S, Sojoudi S (2018) Large-scale sparse inverse covariance estimation via thresholding and max-det matrix completion. *Proc. 35th Internat. Conf. Machine Learn.* (PMLR, *Stockholm, Sweden*), 5766–5775.

Zou H (2006) The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* 101(476):1418–1429.