Advanced Computation of a Sparse Precision Matrix
HADAP: A Hadamard-Dantzig Estimation of a Sparse Precision Matrix

Mohammed Elanbari*, Reda Rawi†, Michele Ceccarelli†, Othmane Bouhali‡, Halima Bensmail†
*Sidra Medical and Research Center, Qatar Foundation
†Qatar Computing Research Institute, Qatar Foundation
‡Texas A&M University-Qatar,
Doha, Qatar

Email: *melanbari@sidra.org; †{rrawi, mceccarelli, hbensmail}@qf.org.qa; ‡othmane.bouhali@qatar.tamu.edu

***Abstract*—Estimating large sparse precision matrices is an interesting and challenging problem in many fields of sciences, engineering, and humanities, thanks to advances in computing technologies. Recent applications often encounter high dimensionality with a limited number of data points leading to a number of covariance parameters that greatly exceeds the number of observations. Several methods have been proposed to deal with this problem, but there is no guarantee that the obtained estimator is positive definite. Furthermore, in many cases, one needs to capture some additional information on the setting of the problem. In this work, we propose an innovative approach named HADAP for estimating the precision matrix by minimizing a criterion combining a relaxation of the gradient-log likelihood and a penalization of lasso type. We derive an efficient Alternating Direction Method of multipliers algorithm to obtain the optimal solution.**

Keywords–Covariance matrix; Frobenius norm; Gaussian graphical model; Precision matrix; Alternating method of multipliers; Positive-definite estimation; Sparsity..

I. INTRODUCTION

Recent applications often encounter high dimensionality with a limited number of data points leading to a number of covariance parameters that greatly exceeds the number of observations. Examples include marketing, e-commerce, and warehouse data in business; microarray, and proteomics data in genomics and heath sciences; and biomedical imaging, functional magnetic resonance imaging, tomography, signal processing, high-resolution imaging, and functional and longitudinal data. In biological sciences, one may want to classify diseases and predict clinical outcomes using microarray gene expression or proteomics data, in which hundreds of thousands of expression levels are potential covariates, but there are typically only tens or hundreds of subjects. Hundreds of thousands of single-nucleotide polymorphisms are potential predictors in genome-wide association studies. The dimensionality of the variables spaces grows rapidly when interactions of such predictors are considered. Large-scale data analysis is also a common feature of many problems in machine learning, such as text and document classification and computer vision. For a $p \times p$ covariance matrix Σ, there are $p(p+1)/2$ parameters to estimate, yet the sample size n is often small. In addition, the positive-definiteness of Σ makes the problem even more complicated. When $n > p$, the sample covariance matrix is positive-definite and unbiased, but as the dimension p increases, the sample covariance matrix tends to become unstable and can fail to be consistent.

II. BACKGROUND

In this paper, we use the notation in Table I.

TABLE I. NOTATION USED IN THIS PAPER

Notation	Description
$\boldsymbol{A} \succeq 0$	$\boldsymbol{A} \in \mathbb{R}^{n\times p}$ is symmetric and positive semidefinite
$\boldsymbol{A} \succ 0$	$\boldsymbol{A} \in \mathbb{R}^{n\times p}$ is symmetric and positive definite
$\|\boldsymbol{A}\|_1$	ℓ_1 norm of $\boldsymbol{A} \in \mathbb{R}^{n\times p}$, i.e $\sum_{ij} a_{ij}$
$\|\boldsymbol{A}\|_\infty$	ℓ_∞ norm of $\boldsymbol{A} \in \mathbb{R}^{n\times p}$, i.e $max_{ij} a_{ij}$
$\|\boldsymbol{A}\|_2$	spectral norm of $\boldsymbol{A} \in \mathbb{R}^{i\times j}$ i.e. the maxmum eigenvalues of $\boldsymbol{A} \succ 0$
$\|\boldsymbol{A}\|_F$	Frobenius norm of $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ i.e $\sqrt{\sum_{ij} a_{ij}^2}$
$\mathrm{Tr}(\boldsymbol{A})$	trace of $\boldsymbol{A} \in \mathbb{R}^{p\times p}$, i.e $\mathrm{Tr}(\boldsymbol{A}) = \sum_i a_{ii}$
vec(A)	stacked form of $\boldsymbol{A} \in \mathbb{R}^{n\times p}$, i.e $\mathrm{vec}(\boldsymbol{A}) = (a_{1,1}, \ldots, a_{n,1}, a_{1,2}, \ldots, a_{1,p}, \ldots, a_{n,p})^t$,
$\mathrm{vec}(\boldsymbol{ABC})$	$(\boldsymbol{C}^t \otimes \boldsymbol{A})\ \mathrm{vec}(\boldsymbol{B})$
$\mathrm{vec}(\boldsymbol{A} \circ \boldsymbol{B})$	$\mathrm{diag}(\mathrm{vec}(\boldsymbol{A}))\mathrm{vec}(\boldsymbol{B})$
$\mathrm{diag}(A)$	$\mathrm{diag}(\boldsymbol{u}) = \begin{pmatrix} u_1 & 0 & \cdots & 0 \\ 0 & u_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_k \end{pmatrix} \in \mathbb{R}^{N\times N}$,
$\boldsymbol{A} \circ \boldsymbol{B}$	Hadamard product of $\boldsymbol{A}$ and $\boldsymbol{B}$, $\in \mathbb{R}^{N\times M}$, i.e element-wise multiplication $(\boldsymbol{A} \circ \boldsymbol{B})_{ij} = (\boldsymbol{A})_{ij} \times (\boldsymbol{B})_{ij}$
$\boldsymbol{A} \otimes \boldsymbol{B}$	Kronecker product of $\boldsymbol{A}$ and $\boldsymbol{B}$ $\boldsymbol{A} \otimes \boldsymbol{B} = \begin{pmatrix} a_{1,1}\boldsymbol{B} & a_{1,2}\boldsymbol{B} & \cdots & a_{1,m}\boldsymbol{B} \\ a_{2,1}\boldsymbol{B} & a_{2,2}\boldsymbol{B} & \cdots & a_{2,m}\boldsymbol{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1}\boldsymbol{B} & a_{n,2}\boldsymbol{B} & \cdots & a_{n,m}\boldsymbol{B} \end{pmatrix}$ $\boldsymbol{A} \otimes \boldsymbol{B} \in \mathbb{R}^{np\times mq}$

a) Existing Methods: Due in part to its importance, there has been an active line of work on efficient optimization methods for solving the ℓ_1 regularized Gaussian MLE problem: PSM that uses projected subgradients [1], ALM using alternating linearization [2], IPM an inexact interior point method [11], SINCO a greedy coordinate descent method [3] and Glass a block coordinate descent method [4][5] etc. For typical high-dimensional statistical problems, optimization methods typically suffer sub-linear rates of convergence [6]. This would be too expensive for the Gaussian MLE problem, since the number of matrix entries scales quadratically with the number of nodes.

b) Sparse Modeling: Sparse modeling has been widely used to deal with high dimensionality. The main assumption is that the p-dimensional parameter vector is sparse, with

 ISBN: 978-1-61208-394-0

many components being exactly zero or negligibly small. Such an assumption is crucial in identifiability, especially for the relatively small sample size. Although the notion of sparsity gives rise to biased estimation in general, it has proved to be effective in many applications. In particular, variable selection can increase the estimation accuracy by effectively identifying important predictors and can improve the model interpretability. To solve this, constraints are frequently imposed on the covariance to reduce the number of parameters in the model including the spectral decomposition, Bayesian methods, modeling the matrix-logarithm, nonparametric smoothing, banding/thresholding techniques (see [5], [7][8]).

Specifically, thresholding is proposed for estimating permutation-invariant consistent covariance matrices when the true covariance matrix is bandable [4]. In this sense, thresholding is more robust than banding/tapering for real applications. In this paper we focus on the soft-thresholding technique as in [4] and [9] because it can be formulated as the solution of a convex optimization problem. In fact, Graphical Lasso approach (Glasso) is introduced as the following. Let $\|.\|_F$ be the Frobenius norm and $|.|_1$ be the element-wise ℓ_1-norm of all non-diagonal elements. Then the soft-thresholding covariance estimator is equal to

$$\hat{\Omega}^{+} = \arg\min_{\Omega} \frac{1}{2}\|\Omega - \hat{\Omega}_n\|_F^2 + \lambda|\Omega|_1, \qquad (1)$$

where $\Omega = \Sigma^{-1}$, where $\Omega = \omega_{ij}{}_{1\leq i,j\leq p}$ is the precision matrix, $\hat{\Omega}$ is the solution of (1) and λ is a tuning parameter. This equation emphasizes the fact that the solution Ω may not be unique [such nonuniqueness can occur if rank$(\Omega) < p$]. It has been shown that the existence of a robust optimization formulation is related to kernel density estimation [10] where property of the solution and a proof that Lasso is consistent was given using robustness directly. Moreover, [11] proved that there exists a linear subspace that is almost surely unique, meaning that it will be the same under different boundary sets corresponding to different solutions of equations of type (1). However, there is no guarantee that the thresholding estimator is always positive definite (see [9], [12] and [13]). Although the positive definite property is guaranteed in the asymptotic setting with high probability, the actual estimator can be an indefinite matrix, especially in real data analysis.

Structure of the inverse covariance matrix has attracted special interest. Example, when dealing with asset allocation in finance. The financial problems are often written in terms of the inverse covariance matrix, in such case the covariance structure of return series is often estimated using only the most recent data, resulting in a small sample size compared to the number of parameters to be estimated. In biological applications (graphical models), zero correlations represent conditional independence between the variables. For example to account for network information in the analysis of metabolites data, the reconstruction of metabolic networks from a collection of metabolite patterns is a key question in the computational research field. Previous attempts focused on linear metabolite associations measured by Pearson correlation coefficients [14]. A major drawback of correlation networks, however, is their inability to distinguish between direct and indirect associations. Correlation coefficients are generally high in large-scale omics data sets, suggesting a plethora of indirect and systemic associations. Gaussian Graphical models (GGMs) circumvent indirect association effects by evaluating conditional dependencies in multivariate Gaussian distributions or equivalently the inverse covariance matrix (see [15]).
On the other hand, additional structure on the precision matrix coefficients is also often required (Comparative genomic hybridizaton) where the difference between two successive coefficients is required to be small or to vary slowly.

Our Contributions: In this paper, we emphasize on introducing a new criteria that insures the positive-definiteness of the covariance matrix adding a tuning parameter $\epsilon > 0$ in the constraints. This additional constraint will guard against positive semi-definite. We add structure on the coefficient of the precision matrix and we derive an efficient ADMM algorithm form to obtain an optimal solution. We perform Alternating Direction Method of Multipliers steps, a variant of the standard Augmented Lagrangian method, that uses partial updates, but with three innovations that enable finessing the caveats detailed above.

In Section III, we link the gaussian graphical model to the precision matrix estimation, in which we show that recovering the structure of a graph G is equivalent to the estimation of the support of the precision matrix and describe different penalized optimization algorithm that have been used to solve this problem and their limitations.

We describe, in Section IV, the ADMM and its application to solve the estimation of the precision matrix under the Dantzig-selector setting and we show its ability to perform distributed optimization where we break up the big optimization problem into smaller problems that are more manageable. In fact, as in the recent methods [9][12], we build on the observation that the Newton direction computation is a Lasso problem, and perform iterative coordinate descent to solve this Lasso problem. Then, we use a Dantzig-selector rule to obtain a step-size that ensures positive-definiteness of the next iterate. In Section V, we validate our algorithm on artificial and real data.

III. Link with Gaussian graphical model (GGM)

Given a data set consisting of samples from a zero mean Gaussian distribution in $\mathbb{R}^p$,

$$X^{(i)} \sim \mathcal{N}(\mathbf{0}, \Sigma), i = 1, ..., n, \qquad (2)$$

with positive definite $p \times p$ covariance matrix Σ. Note that the zero mean assumption in equation 2 is mainly for simplifying notation. The task here is how to estimate the precision matrix $\Omega = \Sigma^{-1}$ when it is sparse. We are particularly interested by the case of sparse Ω because it is closely linked to the selection of graphical models.

To be more specific, let $G = (V, E)$ be a graph representing conditional independence relations between components of $\mathbf{X} = (X_1, ..., X_p)$.

- The vertex set V has p components $X_1, ..., X_p$;
- The edge set E consists of ordered pairs (i, j) of $V \times V$, where $(i, j) \in E$ if there is an edge between X_i and X_j.

We exclude the edge between two vertexes X_i and X_j if and only if X_i and X_j are independent given $\{X_k, k \neq i, j\}$. If in addition $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \Sigma)$, the conditional independence between X_i and X_j given the remaining variables is equivalent

to $\omega_{ij} = 0$, where $\Omega = \Sigma^{-1} = \{\omega_{ij}\}_{1 \leq i,j \leq p}$. Hence, in the case of Gaussian Graphical Model (GGM), the edges are given by the inverse of covariance matrix. More precisely, recovering the structure of a graph G is equivalent to the estimation of the support of the precision matrix Ω. When $n > p$, one can estimate Σ^{-1} by maximum likelihood, but when $p > n$ this is not possible because the empirical covariance matrix is singular and often performs poorly [16]. Since the ℓ_1-norm is the tighest convex upper bound of the cardinality of a matrix, several ℓ_1-regularization methods have been proposed. Consequently a number of authors have considered minimizing an ℓ_1 penalized log-likelihood (see [4] and [17]). It is also sometimes called Glasso [4] after the algorithm that efficiently computes the solution. It consists of solving the penalized problem

$$\hat{\Omega}_{\text{Glasso}} = \arg\min_{\Omega \succeq \mathbf{0}} \{\langle \hat{\Sigma}_n, \Omega \rangle - \log\det\Omega + \lambda \|\Omega\|_1\}. \quad (3)$$

where $\hat{\Sigma}_n$ is the sample covariance matrix and λ is a tuning parameter. The term $\|\Omega\|_1$ encourages sparseness of the precision matrix. The asymptotic properties of the estimator has been studied in [17]. Lauritzen [13] proposed a constrained l1 minimization proposed a constrained l_1 minimization procedure for estimating sparse precision matrices by solving the optimization problem

$$\text{minimize} \quad \|\Omega\|_1 \quad \text{subject to} \quad \|\hat{\Sigma}_n\Omega - \mathbf{I}\|_\infty \leq \lambda, \quad (4)$$

where λ is a tuning parameter. The authors established the rates of convergence under both the entry-wise l_∞ and the Frobenius norm. In computationally viewpoint, equation 4 can be solved by remarking that it can be decomposed into a series of Dantzig selector problems [18]. This observation is useful in both implementation and technical analysis. Theoretically, the authors prove that the estimator is positive definite with high probability. However, in practice there is no guarantee that the estimator is always positive definite, especially in real data analysis.

IV. Alternating Direction Method of Multipliers for HADAP

We define our algorithm HADAP as a solution to the following problem

$$\hat{\Omega}^+ = \arg\min_{\Omega \succeq \epsilon\mathbf{I}} \frac{1}{2}\|\hat{\Sigma}_n\Omega - \mathbf{I}\|_F^2 + \lambda|\mathbf{D} \circ \Omega|_1, \quad (5)$$

where $\hat{\Sigma}_n$ is the empirical covariance matrix, $\mathbf{D} = (d_{ij})_{1 \leq i,j \leq p}$ is an arbitrary matrix with non-negative elements where $\mathbf{D}$ can take different forms: it can be the matrix of all ones or it can be a matrix with zeros on the diagonal to avoid shrinking diagonal elements of Ω. Furthermore, we can take $\mathbf{D}$ with elements $d_{ij} = 1_{i \neq j}|\hat{\Omega}_{ij}^{\text{init}}|$, where $\hat{\Omega}^{\text{init}}$ is an initial estimator of Ω. The later choice of $\mathbf{D}$ corresponds to precision matrix analogue of the Adaptive Lasso penalty.

We propose an ADMM algorithm to solve the problem (5). To derive ADMM, we will first introduce a new variable Θ and an equality constraint as follows:

$$\begin{aligned}(\hat{\Theta}^+, \hat{\Omega}^+) &= \arg\min_{\Theta,\Omega} \left\{ \frac{1}{2}\|\hat{\Sigma}_n\Omega - \mathbf{I}\|_F^2 + \lambda|\mathbf{D} \circ \Omega|_1 \right. \\ &: \left. \Omega = \Theta, \Theta \succeq \epsilon\mathbf{I} \right\}. \end{aligned} \quad (6)$$

The solution to (6) gives the solution to (5). To deal with the problem (6), we have to minimize its augmented Lagrangian function for some given penalty parameter ρ, i.e.

$$\begin{aligned} L_\rho(\Theta, \Omega, \Lambda) &= \frac{1}{2}\|\hat{\Sigma}_n\Omega - \mathbf{I}\|_F^2 + \lambda|\mathbf{D} \circ \Omega|_1 - \langle \Lambda, \Theta - \Omega \rangle \\ &+ \frac{1}{2\rho}\|\Theta - \Omega\|_F^2, \end{aligned} \quad (7)$$

where Λ is the Lagrange multiplier.

At iteration k, the ADMM algorithm consists of the three steps, namely Θ-Step, Ω-Step and the dual-update step :

$$\Theta^{k+1} := \arg\min_{\Theta \succeq \epsilon\mathbf{I}} L_\rho(\Theta, \Omega^k, \Lambda^k); \quad \Theta\text{-minimization} \quad (8)$$

$$\Omega^{k+1} := \arg\min_{\Omega} L_\rho(\Theta^{k+1}, \Omega, \Lambda^k); \quad \Omega\text{-minimization} \quad (9)$$

$$\Lambda^{k+1} := \Lambda^k - \frac{1}{\rho}(\Theta^{k+1} - \Omega^{k+1}). \quad \text{dual-update} \quad (10)$$

- In the first step of the ADMM algorithm, we fix Ω and Λ and minimize the augmented Lagrangian over Θ.
- In the second step, we fix Θ and Λ and minimize the augmented Lagrangian over Ω.
- Finally, we update the dual variable Λ.

To further simplify the ADMM algorithm, we will derive the closed-form solutions for (8)-(10).

A. The Θ-Step.

Let $\mathbf{A}^+$ be the projection of a matrix $\mathbf{A}$ onto the convex cone $\{\Theta \succeq \epsilon\mathbf{I}\}$. Assume that $\mathbf{A}$ has the eigen-decomposition $\sum_{j=1}^p \lambda_j \mathbf{v}_j \mathbf{v}_j^t$. Then, it is well known that $\mathbf{A}^+ = \sum_{j=1}^p \max(\epsilon, \lambda_j)\mathbf{v}_j\mathbf{v}_j^t$. Using this property, the Θ-Step can be analytically solved as follows

$$\begin{aligned} \Theta^{k+1} &= \arg\min_{\Theta \succeq \epsilon\mathbf{I}} L_\rho(\Theta, \Omega^k, \Lambda^k) \\ &= \arg\min_{\Theta \succeq \epsilon\mathbf{I}} \frac{1}{2}\|\hat{\Sigma}_n\Omega^k - \mathbf{I}\|_F^2 + \lambda|\mathbf{D} \circ \Omega^k|_1 \\ &- \langle \Lambda^k, \Theta - \Omega^k \rangle + \frac{1}{2\rho}\|\Theta - \Omega^k\|_F^2 \\ &= \arg\min_{\Theta \succeq \epsilon\mathbf{I}} -\langle \Lambda^k, \Theta \rangle + \frac{1}{2\rho}\|\Theta - \Omega^k\|_F^2 \\ &= \arg\min_{\Theta \succeq \epsilon\mathbf{I}} \|\Theta - (\Omega^k + \rho\Lambda^k)\|_F^2 \\ &= (\Omega^k + \rho\Lambda^k)^+ \\ &= \sum_{j=1}^p \max(\epsilon, \lambda_j)\mathbf{v}_j\mathbf{v}_j^t, \end{aligned}$$

where $\sum_{j=1}^p \lambda_j \mathbf{v}_j \mathbf{v}_j^t$ is eigen-decomposition of $\Omega^k + \rho\Lambda^k$.

 ISBN: 978-1-61208-394-0

B. The Ω-Step.

$$\begin{aligned}
\Omega^{k+1} &= \arg\min_{\Omega} L_\rho(\Theta^{k+1},\Omega,\Lambda^k)\\
&= \arg\min_{\Omega} \frac{1}{2}\|\hat{\Sigma}_n\Omega-\mathbf{I}\|_F^2+\lambda|\mathbf{D}\circ\Omega|_1\\
&- \langle\Lambda^k,\Theta^{k+1}-\Omega\rangle+\frac{1}{2\rho}\|\Theta^{k+1}-\Omega\|_F^2\\
&= \arg\min_{\Omega} \frac{1}{2}\|\hat{\Sigma}_n\Omega-\mathbf{I}\|_F^2+\lambda|\mathbf{D}\circ\Omega|_1\\
&+ \langle\Lambda^k,\Omega\rangle+\frac{1}{2\rho}\|\Theta^{k+1}-\Omega\|_F^2\\
&= \arg\min_{\Omega} \frac{1}{2}\|\hat{\Sigma}_n\Omega-\mathbf{I}\|_F^2+\frac{1}{2\rho}\|\Omega-\Theta^{k+1}\|_F^2\\
&+ \langle\Lambda^k,\Omega\rangle+\lambda|\mathbf{D}\circ\Omega|_1.
\end{aligned}$$

We have

$$\begin{aligned}
\frac{1}{2}\|\hat{\Sigma}_n\Omega-\mathbf{I}\|_F^2 &= \frac{1}{2}\langle\hat{\Sigma}_n\Omega-\mathbf{I},\hat{\Sigma}_n\Omega-\mathbf{I}\rangle\\
&= \frac{1}{2}\left\{\|\hat{\Sigma}_n\Omega\|_F^2-2\langle\hat{\Sigma}_n\Omega,\mathbf{I}\rangle+\|\mathbf{I}\|_F^2\right\}\\
&= \frac{1}{2}\left\{\|\hat{\Sigma}_n\Omega\|_F^2-2\,\mathrm{Tr}(\Omega^t\hat{\Sigma}_n^t\mathbf{I})+\mathrm{Tr}(\mathbf{I}^t\mathbf{I})\right\}\\
&= \frac{1}{2}\left\{\|\hat{\Sigma}_n\Omega\|_F^2-2\,\mathrm{Tr}(\Omega^t\hat{\Sigma}_n)+p\right\}\\
&= \frac{1}{2}\left\{\|\hat{\Sigma}_n\Omega\|_F^2-2\langle\Omega,\hat{\Sigma}_n\rangle+p\right\}.
\end{aligned}$$

and

$$\begin{aligned}
\frac{1}{2\rho}\|\Omega-\Theta^{k+1}\|_F^2 &= \frac{1}{2\rho}\langle\Omega-\Theta^{k+1},\Omega-\Theta^{k+1}\rangle\\
&= \frac{1}{2\rho}\left\{\|\Omega\|_F^2-2\langle\Omega,\Theta^{k+1}\rangle+\|\Theta^{k+1}\|_F^2\right\}.
\end{aligned}$$

Then, the Ω-Step is equivalent to

$$\begin{aligned}
\Omega^{k+1} &= \arg\min_{\Omega} \frac{1}{2}\|\hat{\Sigma}_n\Omega-\mathbf{I}\|_F^2+\frac{1}{2\rho}\|\Omega-\Theta^{k+1}\|_F^2\\
&+ \langle\Lambda^k,\Omega\rangle+\lambda|\mathbf{D}\circ\Omega|_1\\
&= \arg\min_{\Omega} \frac{1}{2}\left\{\|\hat{\Sigma}_n\Omega\|_F^2-2\langle\Omega,\hat{\Sigma}_n\rangle+p\right\}\\
&+ \frac{1}{2\rho}\left\{\|\Omega\|_F^2-2\langle\Omega,\Theta^{k+1}\rangle+\|\Theta^{k+1}\|_F^2\right\}+\langle\Lambda^k,\Omega\rangle\\
&+ \lambda|\mathbf{D}\circ\Omega|_1\\
&= \arg\min_{\Omega} \frac{1}{2}\left\{\|\hat{\Sigma}_n\Omega\|_F^2-2\langle\Omega,\hat{\Sigma}_n\rangle\right\}\\
&+ \frac{1}{2\rho}\left\{\|\Omega\|_F^2-2\langle\Omega,\Theta^{k+1}\rangle\right\}\\
&+ \langle\Lambda^k,\Omega\rangle+\lambda|\mathbf{D}\circ\Omega|_1\\
&= \arg\min_{\Omega} \frac{1}{2}\left\{\|\hat{\Sigma}_n\Omega\|_F^2+\frac{1}{\rho}\|\Omega\|_F^2\right.\\
&- \left.2\langle\Omega,\hat{\Sigma}_n+\frac{1}{\rho}\Theta^{k+1}-\Lambda^k\rangle\right\}+\lambda|\mathbf{D}\circ\Omega|_1\\
&= \arg\min_{\Omega} \frac{1}{2}\left\{\|\hat{\Sigma}_n\Omega\|_F^2+\frac{1}{\rho}\|\Omega\|_F^2\right.\\
&- \left.\frac{2}{\rho}\langle\Omega,\rho(\hat{\Sigma}_n-\Lambda^k)+\Theta^{k+1}\rangle\right\}+\lambda|\mathbf{D}\circ\Omega|_1.
\end{aligned}$$

At this level, we are not able to derive a closed form for Ω. To overcome this problem, we propose to derive a new ADMM to update Ω. To do this, we reparametrize the $\mathbf{D}\circ\Omega$ with Γ and we add an equality constraint $\mathbf{D}\circ\Omega=\Gamma$, then we minimize

$$\frac{1}{2}\left\{\|\hat{\Sigma}_n\Omega\|_F^2+\frac{1}{\rho}\|\Omega\|_F^2-\frac{2}{\rho}\langle\Omega,\rho(\hat{\Sigma}_n-\Lambda^k)+\Theta^{k+1}\rangle\right\}+\lambda|\Gamma|_1$$

subject to

$$\mathbf{D}\circ\Omega=\Gamma. \tag{11}$$

The augmented Lagrangian associated to this problem $L_\rho^k(\Omega,\Gamma,\Delta)$ is

$$\begin{aligned}
&\frac{1}{2}\left\{\|\hat{\Sigma}_n\Omega\|_F^2+\frac{1}{\rho}\|\Omega\|_F^2-\frac{2}{\rho}\langle\Omega,\rho(\hat{\Sigma}_n-\Lambda^k)+\Theta^{k+1}\rangle\right\}\\
+\;&\lambda|\Gamma|_1-\langle\Delta,\Gamma-\mathbf{D}\circ\Omega\rangle+\frac{1}{2\rho}\|\Gamma-\mathbf{D}\circ\Omega\|_F^2,
\end{aligned} \tag{12}$$

where Δ is the Lagrange multiplier and ρ is the same parameter as in (7).
As before, the ADMM for this problem consists of the following three intermediate steps:

$$\Omega_k^{j+1} := \arg\min_{\Omega} L_\rho^k(\Omega,\Gamma^j,\Delta^j);\quad \Omega\text{-minimization} \tag{13}$$

$$\Gamma^{j+1} := \arg\min_{\Gamma} L_\rho^k(\Omega_k^{j+1},\Gamma,\Delta^j);\quad \Gamma\text{-minimization} \tag{14}$$

$$\Delta^{j+1} := \Delta^j-\frac{1}{\rho}(\Omega_k^{j+1}-\Gamma^{j+1}).\quad \text{dual-update} \tag{15}$$

C. The intermediate Ω-Step.

$$\begin{aligned}
\Omega_k^{j+1} &= \arg\min_{\Omega} L_\rho^k(\Omega,\Gamma^j,\Delta^j)\\
&= \arg\min_{\Omega} \frac{1}{2}\left\{\|\hat{\Sigma}_n\Omega\|_F^2+\frac{1}{\rho}\|\Omega\|_F^2\right.\\
&- \left.\frac{2}{\rho}\langle\Omega,\rho(\hat{\Sigma}_n-\Lambda^k)+\Theta^{k+1}\rangle\right\}+\lambda|\Gamma^j|_1\\
&- \langle\Delta^j,\Gamma^j-\mathbf{D}\circ\Omega\rangle+\frac{1}{2\rho}\|\Gamma^j-\mathbf{D}\circ\Omega\|_F^2\\
&= \frac{1}{2}\left\{\|\hat{\Sigma}_n\Omega\|_F^2+\frac{1}{\rho}\|\Omega\|_F^2-\frac{2}{\rho}\langle\Omega,\rho(\hat{\Sigma}_n-\Lambda^k)\right.\\
&+ \left.\Theta^{k+1}\rangle\right\}+\langle\Delta^j,\mathbf{D}\circ\Omega\rangle+\frac{1}{2\rho}\|\Gamma^j-\mathbf{D}\circ\Omega\|_F^2
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2}\,\mathrm{Tr}\left\{\Omega^t\hat{\Sigma}_n^t\hat{\Sigma}_n\Omega+\frac{1}{\rho}\Omega^t\Omega-\frac{2}{\rho}\Omega^t(\rho(\hat{\Sigma}_n-\Lambda^k)+\Theta^{k+1})\right\}\\
&+\mathrm{Tr}((\Delta^j)^t\mathbf{D}\circ\Omega)+\frac{1}{2\rho}\,\mathrm{Tr}((\Gamma^j-\mathbf{D}\circ\Omega)^t(\Gamma^j-\mathbf{D}\circ\Omega)).
\end{aligned}$$

Then, the partial differential of L_ρ^k with respect to Ω, $\frac{\partial L_\rho^k(\Omega,\Gamma^j,\Delta^j)}{\partial\Omega}$ is

$$\begin{aligned}
&= \frac{1}{2}\left\{2\hat{\Sigma}_n^t\hat{\Sigma}_n\Omega+\frac{2}{\rho}\Omega-\frac{2}{\rho}(\rho(\hat{\Sigma}_n-\Lambda^k)+\Theta^{k+1})\right\}\\
&+\Delta^j\circ\mathbf{D}+\frac{1}{2\rho}\left\{-2\Gamma^j\circ\mathbf{D}+2\mathbf{D}\circ\mathbf{D}\circ\Omega\right\}\\
&= \left(\hat{\Sigma}_n^t\hat{\Sigma}_n+\frac{1}{\rho}\mathbf{I}\right)\Omega+\frac{1}{\rho}\mathbf{D}\circ\mathbf{D}\circ\Omega
\end{aligned}$$

$$+\left(\Delta^j-\frac{1}{\rho}\Gamma^j\right)\circ\mathbf{D}-\frac{1}{\rho}\left(\rho(\hat{\Sigma}_n-\Lambda^k)+\Theta^{k+1}\right).$$

Since Ω_k^{j+1} is the minimizer of $L_\rho^k(.,\Gamma^j,\Delta^j)$, we must have

$$\frac{\partial L_\rho^k(\Omega_k^{j+1},\Gamma^j,\Delta^j)}{\partial\Omega} \quad = \quad \mathbf{0},$$

which is equivalent to

$$\left(\hat{\Sigma}_n^t\hat{\Sigma}_n+\frac{1}{\rho}\mathbf{I}\right)\Omega_k^{j+1}+\frac{1}{\rho}\mathbf{D}\circ\mathbf{D}\circ\Omega_k^{j+1}$$

$$+\left(\Delta^j-\frac{1}{\rho}\Gamma^j\right)\circ\mathbf{D}-\frac{1}{\rho}\left(\rho(\hat{\Sigma}_n-\Lambda^k)+\Theta^{k+1}\right)=\mathbf{0}.$$

Finally, Ω_k^{j+1} has a closed form given by the previous expression despite the additional but straightforward computational effort at this level. This additional step, can be considered as a warming start for the original ADMM algorithm. This is very important when dealing with complex problem and large data.

D. The intermediate Γ-Step.

To deal with this Step, define an entry-wise soft-thresholding rule for all the off-diagonal elements of a matrix $\mathbf{A}$ as $\mathbf{S}(\mathbf{A},\kappa)=\{s(a_{jl},\kappa)\}_{1\le j,l\le p}$ with

$$s(a_{jl},\kappa)=\text{sign}(a_{jl})\max(|a_{jl}|-\kappa,0)I_{\{j\neq l\}}.$$

Then the Γ-Step has a closed form given by

$$\begin{aligned}
\Gamma^{j+1} &= \arg\min_\Gamma L_\rho^k(\Omega_k^{j+1},\Gamma,\Delta^j)\\
&= \arg\min_\Gamma \frac{1}{2}\left\{\|\hat{\Sigma}_n\Omega_k^{j+1}\|_F^2+\frac{1}{\rho}\|\Omega_k^{j+1}\|_F^2\right.\\
&- \left.\frac{2}{\rho}\langle\Omega_k^{j+1},\rho(\hat{\Sigma}_n-\Lambda^k)+\Theta^{k+1}\rangle\right\}\\
&+ \lambda|\Gamma|_1-\langle\Delta^j,\Gamma-\mathbf{D}\circ\Omega_k^{j+1}\rangle+\frac{1}{2\rho}\|\Gamma-\mathbf{D}\circ\Omega_k^{j+1}\|_F^2\\
&= \arg\min_\Gamma \frac{1}{2}\|\Gamma-\mathbf{D}\circ\Omega_k^{j+1}\|_F^2-\rho\langle\Delta^j,\Gamma\rangle+\rho\lambda|\Gamma|_1\\
&= \arg\min_\Gamma \frac{1}{2}\left\{\|\Gamma\|_F^2-2\langle\rho\Delta^j+\mathbf{D}\circ\Omega_k^{j+1},\Gamma\rangle\right\}\\
&+ \rho\lambda|\Gamma|_1\\
&= \arg\min_\Gamma \frac{1}{2}\|\Gamma-(\rho\Delta^j+\mathbf{D}\circ\Omega_k^{j+1})\|_F^2+\rho\lambda|\Gamma|_1\\
&= \mathbf{S}(\rho\Delta^j+\mathbf{D}\circ\Omega_k^{j+1},\rho\lambda).
\end{aligned}$$

Algorithm 1 shows complete details of HADAP using ADMM (see Figure 1).

Algorithm 1: HADAP algorithm

initialize the variables: $\Theta^0=\mathbf{0},\Omega^0=\mathbf{0},\Lambda^0=\mathbf{0}$, $\Gamma^0=\mathbf{0}$, $\Delta^0=\mathbf{0}$;
Select a scalar $\rho>0$;
for $k=0,1,2,...$ *until convergence* **do**
  $\Theta^{k+1}:=(\Omega^k+\rho\Lambda^k)^+$;
  for $j=0,1,2,...$ *until convergence* **do**
    $A_\Sigma^{j+1}\leftarrow\left(\rho\hat{\Sigma}_n^t\hat{\Sigma}_n+\mathbf{D}^t\mathbf{D}+\mathbf{I}\right)^{-1}$;
    $B_\Sigma^{j+1}\leftarrow$
    $\left((1-\rho)\mathbf{D}^t\Delta^j+\rho(\hat{\Sigma}_n-\Lambda^k)+\Theta^{k+1}\right)$;
    $\Omega_k^{j+1}:=A_\Sigma\times B_\Sigma$;
    $\Gamma^{j+1}:=\mathbf{S}(\rho\Delta^j+\mathbf{D}\Omega_k^{j+1},\rho\lambda)$;
    $\Delta^{j+1}:=\Delta^j-\frac{1}{\rho}(\Omega_k^{j+1}-\Gamma^{j+1})$;
  end
  $\Omega^{k+1}:=\lim_{j\to+\infty}\Omega_k^j$;
  $\Lambda^{k+1}:=\Lambda^k-\frac{1}{\rho}(\Theta^{k+1}-\Omega^{k+1})$;
end

Figure 1. Complete details of HADAP using the Alternating Direction Method of Multipliers.

V. EXPERIMENTS

Among existing methods, the lasso penalized Gaussian likelihood estimator is the only popular matrix precision estimator that can simultaneously retain sparsity and positive definiteness. To show the goodness of our approach, we use simulations and real example to compare the performance of our estimator with Glasso.

A. Validation on synthetic data

In order to validate our approach, we used the same simulation structure as in [13]. We generated n = 1000 samples from a p = 600-dimensional normal distribution with correlation structure of the form $\sigma(x_i,x_j)=0.6|i-j|$. This model has a banded structure, and the values of the entries decay exponentially as they move away from the diagonal. We generated an independent sample of size 1000 from the same distribution for validating the tuning parameter λ. Using the training data, we compute a series of estimators with 50 different values of λ and use the one with the smallest likelihood loss on the validation sample, where the likelihood loss [19], is defined by

$$L(\Sigma,\Omega)=\langle\Sigma,\Omega\rangle-\log det(\Omega) \tag{16}$$

We mention that all the experiments are conducted on a PC with 4 Gb RAM, 3Ghz CPU using Matlab 2009a.

B. Measurable quantities and results.

Output displays the primal residual $|r^k|$, the primal feasibility tolerance epsilon ϵ^{pri}, the dual residual s^k, and the dual feasibility tolerance ϵ^{dual} quantities. Also included is a plot of the objective values by iterations. Note that the objective value at any particular iteration can go below the true solution value p^* because the iterates does not need to be feasible (e.g., if the constraint is $x-z=0$, we can have $x^k-z^k\neq 0$ for some k).
Results for $\lambda=0.01$ are summarized in Figure 2 and Table II. The convergence is achieved in 25 steps and need just 0.54 seconds. After a few steps of fluctuations ($\approx$ 12 iterations), the objective function stabilizes and converges to its optimal value where the eigenvalues of the precision matrix estimated by the HADAP are real and positive, which prove the positive definiteness of the obtained precision matrix as shown in Figure 3.

 ISBN: 978-1-61208-394-0

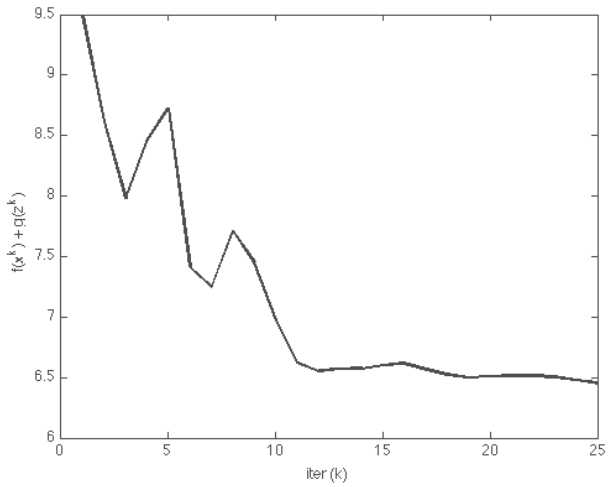

Figure 2. Plot of the objective function for $\lambda = 0.01$

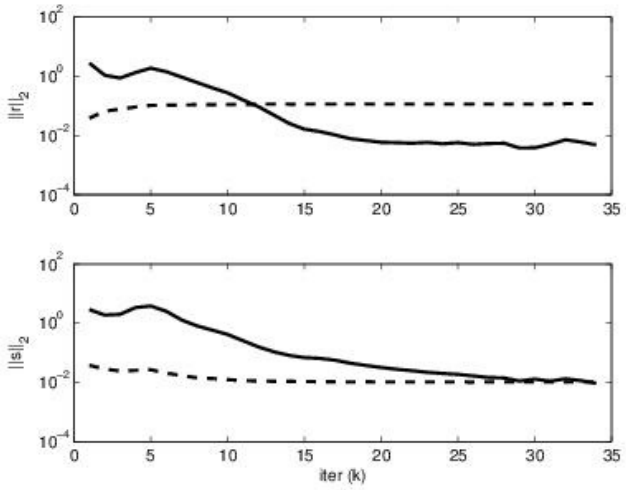

Figure 3. Plot of the objective function

TABLE II. VALUES OF THE PRIMAL RESIDUAL $|r^k|$, THE PRIMAL FEASIBILITY TOLERANCE ϵ^{pri}, THE DUAL RESIDUAL s^k, AND THE DUAL FEASIBILITY TOLERANCE ϵ^{dual} for $\lambda = 0.01$. ALSO INCLUDED THE OBJECTIVE VALUES BY ITERATION.

iter	r.norm	eps.pri	s.norm	eps.dual	objective
1	2.76	0.04	2.86	0.04	8.49
2	1.08	0.07	1.86	0.03	8.11
3	0.88	0.08	1.99	0.02	8.89
4	1.34	0.09	3.37	0.03	7.95
5	1.88	0.10	3.77	0.03	8.64
6	1.42	0.11	2.52	0.02	8.94
7	0.94	0.11	1.28	0.02	7.79
8	0.62	0.11	0.81	0.01	9.70
9	0.41	0.11	0.58	0.01	7.70
10	0.28	0.11	0.42	0.01	6.70
11	0.16	0.11	0.25	0.01	6.71
12	0.09	0.11	0.16	0.01	6.73
13	0.05	0.11	0.11	0.01	6.75
14	0.03	0.11	0.08	0.01	6.77
15	0.02	0.11	0.07	0.01	6.79
16	0.01	0.12	0.06	0.01	6.80
17	0.01	0.12	0.06	0.01	6.82
18	0.01	0.12	0.04	0.01	6.83
19	0.01	0.12	0.04	0.01	6.84
20	0.01	0.12	0.03	0.01	6.85
21	0.01	0.12	0.03	0.01	6.86
22	0.01	0.12	0.02	0.01	6.86
23	0.01	0.12	0.02	0.01	6.87
24	0.01	0.12	0.02	0.01	6.88
25	0.01	0.12	0.02	0.01	6.88

C. Validation on real data

For experimental validation, we used 4 cancer datasets publicly available at the Gene Expression Omnibus [20]. For a fair comparison with the other method of estimating the inverse covariance matrix, we follow the same analysis scheme used by [19]. Datasets are: Liver cancer (GSE1898), Colon cancer (GSE29638), Breast cancer (GSE20194) and Prostate cancer (GSE17951) with sample size $n = 182; 50; 278$ and 154 respectively and number of genes $p = 21794; 22011; 22283$ and 54675. We preprocessed the data so that each variable is zero mean and unit variance across the dataset. We performed 100 repetitions on a $50\% - 50\%$ validation and testing samples.

Since regular sparseness promoting methods do not scale to large number of variables, we used the same regime proposed by [19] and validated our method in two regimes. In the first regime, for each of the 50 repetitions, we selected $n = 200$ variables uniformly at random and use the glasso. In the second regime, we use all the variables in the dataset, and use the method dpglasso from [21]. Since the whole sample covariance matrix could not fit in memory, we computed it in batches of rows [21]. In order to make a fair comparison, the runtime includes the time needed to produce the optimal precision matrix from a given input dataset. Average runtimes was summarized in table III. This includes not only the time to solve each optimization problem but also the time to compute the covariance matrix (if needed). Our HADAP method is considerably faster than the Glasso method as shown in table III .

TABLE III. RUNTIMES FOR GENE EXPRESSION DATASETS. OUR HADAP METHOD IS CONSIDERABLY FASTER THAN SPARSE METHOD.

Dataset	*Graphical lasso*	*Our estimator*
GSE1898	3.8 min	1.0 min
GSE29638	3.8 min	2.6 min
GSE20194	3.8 min	2.5 min
GSE17951	14.9min	4.8 min

VI. Conclusion and Future Work

The sparse precision matrix estimator has been shown to be useful in many applications. Penalizing the matrix is a tool with good asymptotic properties for estimating large sparse covariance and precision matrices. However, its positive definiteness property and unconstrained structure can be easily violated in practice, which prevents its use in many important applications such as graphical models, financial assets and comparative genomic hybridization. In this paper, we have expressed the precision matrix estimation equation in a convex optimization framework and considered a natural modification by imposing the positive definiteness and problem-solving constraints. We have developed a fast alternating direction method to solve the constrained optimization problem and the resulting estimator retains the sparsity and positive definiteness properties simultaneously. We are at the phase of demonstrating the general validity of the method and its advantages over

 ISBN: 978-1-61208-394-0

correlation networks based on competitive precision matrix estimators with computer-simulated reaction systems, to be able to demonstrate strong signatures of intracellular pathways and provide a valuable tool for the unbiased reconstruction of metabolic reactions from large-scale metabolomics data sets.

REFERENCES

[1] J. Duchi, S. Gould, and D. Koller, "Projected Subgradient Methods for Learning Sparse Gaussians," in Proceedings of the Twenty-fourth Conference on Uncertainty in AI (UAI), 2008.

[2] K. Scheinberg, S. Ma, and D. Goldfarb, "Sparse inverse covariance selection via alternating linearization methods," in Advances in Neural Information Processing, NIPS10, 2010.

[3] L. Li and K. Toh, "An inexact interior point method for l1-reguarlized sparse covariance selection," Mathematical Programming Computation, vol. 2, 2010, pp. 291–315.

[4] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse Inverse Covariance Estimation With the Graphical Lasso," Biostatistics, vol. 9, 2008, pp. 432–441.

[5] O. Banerjee, L. El Ghaoui, and A. d'Aspremont, "Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data," Journal of Machine Learning Research, vol. 9, 2008, pp. 485–516.

[6] A. Agarwal, S. Negahban, and M. Wainwright, "Convergence rates of gradient methods for high-dimensional statistical recovery," in Advances in Neural Information Processing, NIPS10, 2010.

[7] P. Bickel and E. Levina, "Regularized estimation of large covariance matrices," Ann. Statist., vol. 36, 2008a, pp. 199–227.

[8] R. Yang and J. Berger, "Estimation of a covariance matrix using the reference prior," Ann. Statist., vol. 3, 1994, pp. 1195–1211.

[9] X. Lingzhou, S. Ma, and H. Zhou, "Positive Definite L_1 Penalized Estimation of Large Covariance Matrices," Journal of the American Statistical Association, vol. 500, 2012, pp. 1480–1491.

[10] H. Xu, C. Caramanis, and S. Mannor, "Robust Regression and Lasso," IEEE Transaction on Information Theory, vol. 56, 2010, pp. 3561–357.

[11] R. Tibshirani and J. Taylor, "The solution path of the generalized lasso," Ann. Statist, vol. 39 (3), 2011, pp. 1335–1371.

[12] N. El Karoui, "Operator norm consistent estimation of large dimensional sparse covariance matrices," Annals of Statistics, vol. 36, 2008, pp. 2717–2756.

[13] T. Cai and H. Zhou, "A constrained l_1 minimization approach to sparse precision matrix estimation," J. American Statistical Association, vol. 106, 2011, pp. 594–607.

[14] J. Krumsiek, K. Suhre, T. Illig, J. Adamski, and F. Theis, "Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data," BMC Systems Biology, vol. 5 (21), 2011, pp. 2–16.

[15] S. Lauritzen, Graphical Models. Clarendon Press, Oxford, 1996.

[16] I. Johnstone, "On the distribution of the largest eigenvalue in principal components analysis," Ann. Statist., vol. 29 (2), 2001, pp. 295–327.

[17] P. Ravikumar, W. M.J., G. Raskutti, and B. Yu, "High-dimensional covariance estimation by minimizing l_1-penalized log-determinant divergence," Electron. J. Statist., vol. 5, 2011, pp. 935–980.

[18] E. Candes and T. Tao, "The Dantzig selector: statistical estimation when p is much larger than n," Annals of Statistics, vol. 35, 2007, pp. 2313–2351.

[19] J. Honorio and T. Jaakkola, "Inverse covariance estimation for high-dimensional data in linear time and space: Spectral methods for riccati and sparse models," in Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence, 2013.

[20] R. Edgar, M. Domrachev, and A. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," Nucleic Acids Res, vol. 30 (1), 2002, pp. 207–210.

[21] R. Mazumder and T. Hastie, "Exact covariance thresholding into connected components for largescale graphical lasso," The Journal of Machine Learning Research, vol. 13, 2012, pp. 781–794.

 ISBN: 978-1-61208-394-0