Wuhan University
Journal of Natural Sciences

2022, Vol.27 No.2, 128-134

Article ID 1007-1202(2022)02-0128-07
DOI https://doi.org/10.1051/wujns/2022272128

ADMM Algorithmic Regularization Paths for Sparse and Large Scale Positive-Definite Covariance Matrix Estimation

☐ XIA Lin[1], WANG Guanpeng[2], HUANG Xudong[3]

1. School of Computer and Software Engineering, Anhui Institute of Information Technology, Wuhu 241002, Anhui, China;

2. School of Mathematical Sciences, Capital Normal University, Beijing 100048, China;

3. School of Mathematics and Statistics, Anhui Normal University, Wuhu 241002, Anhui, China

Abstract: Estimating sparse positive-definite covariance matrices in high dimensions has received extensive attention in the past two decades. However, many existing algorithms are proposed for a single regularization parameter and little attention has been paid to estimating the covariance matrices over the full range of regularization parameters. In this paper we suggest to compute the regularization paths of estimating the positive-definite covariance matrices through a one-step approximation of the warm-starting Alternating Direction Method of Multipliers (ADMM) algorithm, which quickly outlines a sequence of sparse solutions at a fine resolution. We demonstrate the effectiveness and computational savings of our proposed algorithm through elaborative analysis of simulated examples.

Key words: alternating direction method of multiplier; covariance matrix; high dimensionality; regularization parameters; sparse estimation

CLC number: O 212

Received date: 2021-11-04

Foundation item: Supported by the Natural Science Research Project of Anhui Universities (KJ2020A0823), Department of Science and Technology of Anhui Province General Project(1908085MA20), Provincial Quality Project of Anhui (2019JYXM0894), and the Demonstration and Leading Base of First-Class Undergraduate Talents in Software Engineering (2019RCSFJD100)

Biography: XIA Lin, male, Lecturer, research direction: high dimensional statistical inference. E-mail: 1491357861@qq.com

0 Introduction

High dimensional data can be efficiently collected at a low cost in many scientific areas due to the advance of information technology. In analysis of high dimensional data, estimating the positive-definite covariance matrix, denoted $\boldsymbol{\Sigma}$, is perhaps one of the most fundamental problems. It is however not straightforward to estimate $\boldsymbol{\Sigma}$ precisely in high dimensions. To estimate the positive-definite covariance matrix $\boldsymbol{\Sigma}$ accurately, an important concept, sparsity, is introduced which assumes many off-diagonal components of $\boldsymbol{\Sigma}=(\sigma_{i,j})_{p\times p}$ are identically zero. The sparsity assumption plays an important role to construct $\hat{\boldsymbol{\Sigma}}=(\sigma_{i,j})_{p\times p}$, an intuitively feasible estimate of $\boldsymbol{\Sigma}$, and to establish the consistency property of $\hat{\boldsymbol{\Sigma}}$. If $\boldsymbol{\Sigma}$ is sparse, many interesting approaches, such as banding[1], tapering[2] and thresholding[1,3] are proposed in the literature to produce sparse estimates of $\boldsymbol{\Sigma}$. However, these sparse estimates are not always positive-definite, though the positive-definite property is highly desirable.

The definition of positive-definite was described in Xue *et al*[4]. When the positive-definite property is concerned in estimating the high dimensional covariance matrix, Friedman *et al*[5] and Rothman[6] considered the following minimization problem:

$$\hat{\boldsymbol{\Sigma}}=(\tau,\lambda)=\arg\min_{\boldsymbol{\Sigma}\geq 0}\left\{\frac{1}{2}\left\|\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}_n\right\|_2^2-\tau\log\{\det(\boldsymbol{\Sigma})\}+\lambda\|\boldsymbol{\Sigma}\|_1\right\} \tag{1}$$

where τ and λ are two distinctive regularization

parameters, $\tau \geq 0$ is introduced here to ensure the positive-definite of $\hat{\boldsymbol{\Sigma}}=(\tau,\lambda)$ and $\lambda \geq 0$ controls the trade-off between the penalty and the loss function. The following notations will be used:

$\hat{\boldsymbol{\Sigma}}_n=n^{-1}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^{\mathrm{T}}$ is the sample covariance matrix, $\bar{x}=n^{-1}\sum_{i=1}^{n}x_i, x_1,\cdots,x_n$ is a random sample of size $n, \boldsymbol{X}_i=(X_{i,1},\cdots,X_{i,p})^{\mathrm{T}}\in\mathbf{R}^p$, $\det(\boldsymbol{\Sigma})$ stands for the determinant of $\boldsymbol{\Sigma}$, $\|\boldsymbol{\Sigma}\|_2$ stands for the Frobenius norm of $\boldsymbol{\Sigma}$, and $\|\boldsymbol{\Sigma}\|_1$ stands for the ℓ_1 norm of all off-diagonal elements of $\boldsymbol{\Sigma}$. Specifically,

$$\|\boldsymbol{\Sigma}\|_2=\sum_{i,j=1}^{p}\sigma_{i,j}^2 \text{ and } \|\boldsymbol{\Sigma}\|_1=\sum_{i\neq j}|\sigma_{i,j}|$$

An iterative shooting algorithm is proposed by Friedman *et al* [5] and Rothman [6] to solve (1). However, whether this shooting algorithm converges or not remains unknown in the literature. To yield a positive-definite estimate of $\boldsymbol{\Sigma}$, Xue *et al* [4] considered a different optimization problem:

$$\hat{\boldsymbol{\Sigma}}(\lambda)=\arg\min_{\boldsymbol{\Sigma}\geq\varepsilon\boldsymbol{I}}\left\{\frac{1}{2}\|\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}_n\|_2^2+\lambda\|\boldsymbol{\Sigma}\|_1\right\} \quad (2)$$

for a sufficiency small positive constant ε, where $\boldsymbol{I}$ stands for an identity matrix. Solving (2) with the constraint $\boldsymbol{\Sigma}\geq\varepsilon\boldsymbol{I}$ must yield a positive-definite $\hat{\boldsymbol{\Sigma}}(\lambda)$. Xue *et al* [4] proposed an Alternating Direction Method of Multiplies (ADMM) to convert the problem of solving (2) into a sequence of simpler optimization subproblems. This method has been cited and studied in many subsequent researches [4].

Although solving a sequence of subproblems may end up taking more iterations than directly solving (2), it runs in less total time because the subproblems are relatively easy to solve in general. In the present context, all the subproblems have explicit analytic forms. These two optimization problems, described in (1) and (2) can be efficiently solved for a single regularization parameter, referring to Kang and Deng [7], Choi *et al* [8], Zhang *et al* [9], etc. However, when these optimization problems are solved over the full range of regularization parameters, the computational complexity will be dramatically increased.

Efficient estimation of positive-definite covariance matrix can be implemented by these two optimization problems, described in (1) and (2). But Log structure of (1) is not stable in estimation process. In this paper we suggest to compute the regularization paths of $\hat{\boldsymbol{\Sigma}}(\lambda)$ through a slight modification of Xue *et al*'s ADMM algorithm [4]. To choose an optimal λ which yields the "best" estimate $\hat{\boldsymbol{\Sigma}}(\lambda)$ in a certain sense, Xue *et al* [4] suggested to inspect a sequence of sparse solutions over the full range of regularization parameters: $\{\hat{\boldsymbol{\Sigma}}=(\lambda):0\leq\lambda_1\leq\cdots\leq\lambda_M\}$, where M is usually set to 100 heuristically, $\lambda_M=\lambda_{\max}$ and $\lambda_{\max}$ is the maximal value of λ such that $\hat{\boldsymbol{\Sigma}}(\lambda)=0$, the sparsest solution. Conceptually, estimating $\hat{\boldsymbol{\Sigma}}(\lambda)$ at a fixed λ consists of two ingredients. The first is to seek for the support of $\hat{\boldsymbol{\Sigma}}(\lambda)=\{\hat{\sigma}_{i,j}(\lambda)\}_{p\times p}$, which is defined as the (i,j) pairs such that $\hat{\sigma}_{i,j}(\lambda)\neq 0$. The second step is to solve (2) over the support of $\hat{\boldsymbol{\Sigma}}(\lambda)$. In general, the first step is much more important than the second step for large scale problems because one can always refit $\boldsymbol{\Sigma}$ to obtain an optimal estimate of the parameter values with a sufficiently correct estimate of the support of $\boldsymbol{\Sigma}$. To handle large scale problems, we need algorithms not only to recover the support of $\hat{\boldsymbol{\Sigma}}(\lambda)$ but also to estimate the entire sequence of $\hat{\boldsymbol{\Sigma}}(\lambda)$ quickly. Following Friedman *et al* [10], Wu and Lange [11], Hu *et al* [12], we first introduce a procedure to estimate the regularization paths through an ADMM algorithm with warm starts over a range of regularization parameters to yield a path-like sequence of solutions, then perform a one-step approximation along each point on this path to yield an ADMM algorithmic regularization path. This procedure quickly outlines a sequence of $\hat{\boldsymbol{\Sigma}}(\lambda)$ at a fine resolution and, at the same time, reduces computational time dramatically.

The rest of this paper is organized as follows. In Section 1, we first review the ADMM algorithm with warm-starts over a grid of regularization parameters to yield a path-like sequence of solutions, then use the one-step approximation of this path-like sequence of solutions to obtain an algorithmic regularization path. In Section 2, we demonstrate the effectiveness and computational savings of our proposals through simulated examples. We conclude this paper with some brief remarks in Section 3.

1 The Algorithmic Regularization Path

In this section we use Xue *et al* 's [4] ADMM splitting method as the foundation to develop a new approximation to the sequence of sparse solutions $\hat{\boldsymbol{\Sigma}}(\lambda)$, which allows us to quickly explore the space of sparse solutions at a fine resolution. And, sacrifice of estimation

precision can be ignored as long as the step size, defined as $\max_j(\lambda_{j+1}-\lambda_j)$, is sufficiently small. The objective function formulated in (2) by Xue *et al*[4] consists of a differentiable loss function, $\left\|\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}_n\right\|_2^2$, a non-differentiable penalty, $\lambda\|\boldsymbol{\Sigma}\|_1$, and a constraint $(\boldsymbol{\Sigma}\geq\varepsilon\boldsymbol{I})$ which ensures that $\hat{\boldsymbol{\Sigma}}(\lambda)$ is positive-definite. In the constraint, ε is a user-specified parametric which is sufficiently small, say, $\varepsilon=10^{-6}$. The loss function is also convex in $\boldsymbol{\Sigma}$. The regularization parameter $\lambda\geq 0$ controls the trade-off between of the loss function and the penalty. Following Xue *et al* [4], we split the smooth loss function from the nonsmooth penalty through a copy para matrix $\boldsymbol{\Theta}$, and add an equality constraint forcing $\boldsymbol{\Sigma}=\boldsymbol{\Theta}$, which converts (2) into

$$(\hat{\boldsymbol{\Theta}}^+,\hat{\boldsymbol{\Sigma}}^+)=\arg\min_{\boldsymbol{\Sigma},\boldsymbol{\Theta}}\left\{\frac{1}{2}\left\|\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}_n\right\|_2^2+\lambda\|\boldsymbol{\Theta}\|_1:\boldsymbol{\Sigma}=\boldsymbol{\Theta},\boldsymbol{\Theta}\geq\varepsilon\boldsymbol{I}\right\} \tag{3}$$

Similar operator splitting strategies are also used in many other learning problems such as total variation[13], jointly graphical model selection[14,15], a convex formulation of clustering[16], etc. With scaled dual variable $\boldsymbol{\Lambda}$ of the same dimension as $\boldsymbol{\Sigma}$ and $\boldsymbol{\Theta}$ and an algorithm tuning parameter μ the associated augmented Lagrangian of (3) is

$$\phi(\boldsymbol{\Sigma},\boldsymbol{\Theta},\boldsymbol{\Lambda})=\frac{1}{2}\left\|\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}_n\right\|_2^2+\lambda\|\boldsymbol{\Theta}\|_1+\frac{\mu}{2}\left\|\boldsymbol{\Sigma}-\boldsymbol{\Theta}+\boldsymbol{\Lambda}\right\|_2^2$$

Then the ADMM algorithm for (2) is partitioned into three subproblems:

1) $\boldsymbol{\Sigma}$-subproblem:

$$\boldsymbol{\Sigma}^k=\arg\min_{\boldsymbol{\Sigma}}\{\phi(\boldsymbol{\Sigma},\boldsymbol{\Theta}^{k-1},\boldsymbol{\Lambda}^{k-1})\} \tag{4}$$

2) $\boldsymbol{\Theta}$-subproblem:

$$\boldsymbol{\Theta}^k=\arg\min_{\boldsymbol{\Theta}}\{\phi(\boldsymbol{\Sigma}^k,\boldsymbol{\Theta},\boldsymbol{\Lambda}^{k-1})\},\ \text{subject to}\ \boldsymbol{\Theta}\geq\varepsilon\boldsymbol{I} \tag{5}$$

3) Dual update

$$\boldsymbol{\Lambda}^k=\boldsymbol{\Lambda}^{k-1}+\boldsymbol{\Sigma}^k-\boldsymbol{\Theta}^k$$

We solve these subproblems, together with the dual update, iteratively until convergence. The benefit of solving these subproblems is that there are closed-form solutions to both (4) and (5). The $\boldsymbol{\Sigma}$-subproblem solves a linear regression with an additional quadratic ridge penalty. Specifically,

$$\begin{aligned}\boldsymbol{\Sigma}^k&=\arg\min_{\boldsymbol{\Sigma}}\\&\{\frac{1}{2}\left\|\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}_n\right\|_2^2+\lambda\left\|\boldsymbol{\Theta}^{k-1}\right\|_1+\frac{\mu}{2}\left\|\boldsymbol{\Sigma}-\boldsymbol{\Theta}^{k-1}+\boldsymbol{\Lambda}^{k-1}\right\|_2^2\}\\&=\arg\min_{\boldsymbol{\Sigma}}\{\frac{1}{2}\left\|\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}_n\right\|_2^2+\frac{\mu}{2}\left\|\boldsymbol{\Sigma}-\boldsymbol{\Theta}^{k-1}+\boldsymbol{\Lambda}^{k-1}\right\|_2^2\}\\&=(1+\mu)^{-1}\{\mu(\boldsymbol{\Theta}^{k-1}-\boldsymbol{\Lambda}^{k-1})+\hat{\boldsymbol{\Sigma}}_n\}\end{aligned}$$

The $\boldsymbol{\Theta}$-subproblem also has an analytical solution. For notational clarity, we define $(\boldsymbol{B})_+$ as the projection of $(\boldsymbol{B})=(B_{j,l})_{1\leq j,l\leq p}$ onto the convex cone $\{\boldsymbol{\Theta}\geq\varepsilon\boldsymbol{I}\}$. Specifically, suppose λ_j is the j-th eigenvalue of $\boldsymbol{B}$ and μ_j is the associated eigenvector, namely,

$$\boldsymbol{B}=\sum_{j=1}^p\lambda_j\boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\mu}_j,\ \text{then}\ (\boldsymbol{B})_+=\sum_{j=1}^p\lambda_j\max(\lambda_j,\varepsilon)\boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\mu}_j$$

We further define an entry wise soft thresholding rule, denoted $\boldsymbol{S}(\boldsymbol{B},\tau)=\{s(B_{j,l},\tau)_{1\leq j,l\leq p}\}_{1\leq j,l\leq p}$, for all the off-diagonal elements of $\boldsymbol{B}$.

$$\begin{aligned}s(B_{j,l},\tau)&=\operatorname{sign}(B_{j,l})\max(\left|B_{j,l}\right|-\tau,0)\boldsymbol{I}\\&(j\neq l)+B_{j,l}\boldsymbol{I}(j=l)\end{aligned}$$

The analytical solution of the $\boldsymbol{\Theta}$-subproblem is as follows:

$$\begin{aligned}\boldsymbol{\Theta}^k&=\arg\min_{\boldsymbol{\Theta}\geq\varepsilon\boldsymbol{I}}\\&\{\frac{1}{2}\left\|\boldsymbol{\Sigma}^k-\hat{\boldsymbol{\Sigma}}_n\right\|_2^2+\lambda\|\boldsymbol{\Theta}\|_1+\frac{\mu}{2}\left\|\boldsymbol{\Sigma}^k-\boldsymbol{\Theta}+\boldsymbol{\Lambda}^{k-1}\right\|_2^2\}\\&=\arg\min_{\boldsymbol{\Theta}\geq\varepsilon\boldsymbol{I}}\{\lambda\|\boldsymbol{\Theta}\|_1+\frac{\mu}{2}\left\|\boldsymbol{\Theta}-(\boldsymbol{\Sigma}^k+\boldsymbol{\Lambda}^{k-1})\right\|_2^2\}\\&=\{\boldsymbol{S}(\boldsymbol{\Sigma}^k+\boldsymbol{\Lambda}^{k-1},\lambda/\mu)\}_+\end{aligned}$$

Solving the $\boldsymbol{\Theta}$-subproblem introduces sparsity. Finally, the dual update guarantees that $\boldsymbol{\Sigma}^k$ is squeezed towards $\boldsymbol{\Theta}^k$. As the iterations proceed, $\boldsymbol{\Theta}^k$ becomes increasingly sparse. We shall illustrate this phenomenon with simulations in Section 2. One can refer to Fig. 1 (a) and (c) for details. Because the $\boldsymbol{\Theta}$-subproblem controls the sparsity level of $\boldsymbol{\Theta}^k$, a natural question arises: How to use or modify the iterates of the ADMM algorithm to quickly generate regularization paths for estimating a sparse positive-definite covariance matrix in high or even ultrahigh dimensions?

Following the idea of Hu *et al* [12], we provide two options to quickly generate a regularization path. The first is to use warm starts in the ADMM algorithm along a grid of regularization parameters, say, $0\leq\lambda_1\leq\cdots\leq\lambda_M$. To be precise, the warm starts use the solution from the previous value of λ_j, $\hat{\boldsymbol{\Sigma}}(\lambda_j)$, as the initial value for the ADMM algorithm to solve the optimization problem at λ_{j+1}. If λ_{j+1} is sufficiently close to λ_j, it is reasonable to expect that $\hat{\boldsymbol{\Sigma}}(\lambda_{j+1})$ is also very close to $\hat{\boldsymbol{\Sigma}}(\lambda_j)$. Using warm starts usually reduces the number of iterations dramatically.

The ADMM algorithm with warm starts, which is referred to as Algorithm 1 in our context, is described below:

Algorithm 1: ADMM with warm starts to estimate sparse and large scale positive-definite covariance

1. Initialize $\boldsymbol{\Sigma}^0 = 0, \boldsymbol{\Lambda}^0 = 0,$ and M log-spaced values $\lambda_1 \prec \lambda_2 \prec \cdots \prec \lambda_M$ for $\lambda_1 = 0$ and $\lambda_M = \lambda_{\max}$
2. for $j = 1, \cdots, M$ do
   Initialize $\boldsymbol{\Sigma}^0(\lambda_j) = \hat{\boldsymbol{\Sigma}}(\lambda_{j-1})$ and $\boldsymbol{\Lambda}^0(\lambda_j) = \hat{\boldsymbol{\Lambda}}(\lambda_{j-1})$;
   while $\|\boldsymbol{r}^k\| \wedge \|\boldsymbol{s}^k\| > \varepsilon^{\text{tolerance}}$ do
   $\boldsymbol{\Theta}^k(\lambda_j) = \left[\boldsymbol{S}\{\boldsymbol{\Sigma}^{k-1}(\lambda_j) + \boldsymbol{\Lambda}^{k-1}(\lambda_j), \lambda_j / \mu\}\right]_+$
   $\boldsymbol{\Sigma}^k(\lambda_j) = (1+\mu)^{-1}\left[\mu\{\boldsymbol{\Theta}^k(\lambda_j) - \boldsymbol{\Lambda}^{k-1}(\lambda_j)\} + \hat{\boldsymbol{\Sigma}}_n\right]$
   $\boldsymbol{\Lambda}^k(\lambda_j) = \boldsymbol{\Lambda}^{k-1}(\lambda_j) + \boldsymbol{\Sigma}^k(\lambda_j) - \boldsymbol{\Theta}^k(\lambda_j)$
   $\boldsymbol{r}^k = \boldsymbol{\Sigma}^k(\lambda_j) - \boldsymbol{\Theta}^k(\lambda_j)$ and $\boldsymbol{S}^k = \boldsymbol{\Theta}^k(\lambda_j) - \boldsymbol{\Theta}^{k-1}(\lambda_j)$
   $k = k + 1$
   end while
   end for
3. Output $\{\boldsymbol{\Sigma}(\lambda_j), j = 1, \cdots, M\}$ as the regularization path.

In the above algorithm, the convergence is measured by the primal and dual residuals[17]. The resulting solution goes from dense to sparse, or equivalently, the regularization parameter λ goes from small to large. The ADMM algorithm with warm starts can certainly go in the reverse direction from sparse to dense. However, going from sparse to dense may introduce additional yet unnecessary discontinuities in the $\boldsymbol{\Theta}$-subproblem, consequently would require more iterations for convergence. Therefore, we advocate using the ADMM algorithm with warm starts going from dense to sparse, as described in Algorithm 1 at present.

The second option to speed the computation of regularization paths is to seek for a single path approximating algorithm instead of solving several optimization problems separately over a grid of regularization parameter values. In general, the sparsity levels of $\boldsymbol{\Theta}^k(\lambda)$ typically stabilize to that of the solution within a small number of iterations as λ is increased from λ_1 to λ_M; and the remaining iterations and a large proportion of the computation time are spent on squeezing $\boldsymbol{\Sigma}$ towards $\boldsymbol{\Theta}$ to satisfy the primal feasibility. This motivates us to surmise that, if λ_{j+1} is sufficiently close to λ_j, then the sparsity patterns given by the $\boldsymbol{\Theta}$-subproblem may correctly approximate the regularization paths within a few or even one iteration when implementing the ADMM algorithm with warm starts. This further motivates us to suggest the following ADMM Algorithmic Regularization Paths to quickly approximate the regularization paths for estimating large scale positive-definite covariance matrices.

This algorithm is described as follows:

Algorithm 2: Algorithmic regularization paths to estimate sparse and large scale positive-definite covariance matrices

1. Initialize $\boldsymbol{\Theta}^0 = 0, \boldsymbol{\Lambda}^0 = 0,$ set λ to be $\lambda_1 < \lambda_2 < \cdots < \lambda_M$ for $\lambda_1 = 0$ and $\lambda_M = \lambda_{\max}$
2. while $\|\boldsymbol{\Theta}^j\|_{1,\text{off}} \neq 0$ do
   $\boldsymbol{\Sigma}^j(\lambda_j) = (1+\mu)^{-1}\left[\mu\{\boldsymbol{\Theta}^{j-1}(\lambda_{j-1}) - \boldsymbol{\Lambda}^{j-1}(\lambda_{j-1})\} + \hat{\boldsymbol{\Sigma}}_n\right]$
   $\boldsymbol{\Theta}^j(\lambda_j) = \left[\boldsymbol{S}\{\boldsymbol{\Sigma}^j(\lambda_j) + \boldsymbol{\Lambda}^{j-1}(\lambda_{j-1}), \lambda_j / \mu\}\right]_+$
   $\boldsymbol{\Lambda}^j(\lambda_j) = \boldsymbol{\Lambda}^{j-1}(\lambda_{j-1}) + \boldsymbol{\Sigma}^j(\lambda_j) - \boldsymbol{\Theta}^j(\lambda_j)$
   $j = j + 1$
   end while
3. Output $\{\boldsymbol{\Theta}^j(\lambda_j), j = 1, \cdots, M\}$ as the regularization path.

In Algorithm 2, the sequence of λ_j can be linearly spaced. As long as the step size, defined as $\max_j(\lambda_{j+1} - \lambda_j)$, is sufficiently small, the sparsity patterns of Algorithm 2 can well approximate the regularization paths of $\hat{\boldsymbol{\Sigma}}(\lambda)$. This algorithm makes full use of three key ingredients to quickly generate a regularization path: the sparsity patterns of the solution to the $\boldsymbol{\Theta}$-subproblem, the ADMM with warm-starts generating a dense to sparse solution and the one-step approximation at each regularization level. We begin with the fully dense ridge solution, employing these key ingredients, to implement the ADMM algorithm which gradually increases the amount of regularization level until we obtain a fully sparse solution. If we iterate the three subproblems until convergence, Algorithm 2 would be approximatively equivalent to Algorithm 1. Therefore, the one-step approximation is the key difference between these two algorithms.

In both of the above algorithms, there is an ADMM tuning parameter μ. We fix $\mu = 1$ in our algorithms. Updating μ dynamically may speed up convergence (Boyd *et al*[17], He *et al*[18]), however, it is not conducive to achieve a path-like algorithm using warm-starts. In particular, the sparsity levels of $\boldsymbol{\Theta}$ dramatically change when μ is changed because the $\boldsymbol{\Theta}$-subproblem was solved by soft-thresholding at the level λ / μ, which eliminates the benefit of using warm-starts to achieve relatively smooth transitions in sparsity levels.

There are at least three advantages of working with Algorithm 2: it is easy to implement with a few lines of code; it fleetly gives a sequence of sparse solutions as it requires M iterations to fully explore the regularization paths $\hat{\boldsymbol{\Sigma}}(\lambda)$; and it also allows us to explore the space of sparse models at a very fine resolution.

2 Numerical Studies

We illustrate the performance of our proposals through simulations. Write $\boldsymbol{\Sigma}=(\sigma_{i,j})_{p\times p}$. We set the sample size $n=50$ and vary the dimension $p=100$, 200, 500 and 1 000. We consider the following two models which were once used in existing literature:

1) Model (I): $\sigma_{i,j}=(1-|i-j|/10)_+$.

2) Model (II): We partition the index set, $\{1,\cdots,p\}$, into $K=p/20$ non-overlapping subsets of equal size, and denote I_k the index set, $k=1,\cdots,K$. Let

$$\sigma_{i,j}=0.6I(i=j)+0.4\sum_{k=1}^{k}I(i\in I_k,j\in I_k)$$
$$+0.4\sum_{k=1}^{K-1}\{I(i=I_k,j\in I_{k+1})+I(i\in I_{k+1},j=I_k)\}$$

Specifically, Model (I) was used by Cai *et al*[2] and Model (II) was used by Rothman *et al*[5].

Figure 1 shows the sparsity levels. In Fig. 1 (a) and (c) we report the sparsity levels, measured by the number of nonzero components of $\hat{\boldsymbol{\Theta}}$ of the ADMM algorithm with warm starts (Algorithm 1) for a fixed λ when $p=1\,000$. The ADMM algorithm with warm starts spends a large proportion of the computational time on squeezing $\boldsymbol{\Sigma}$ towards $\boldsymbol{\Theta}$ to satisfy the primal feasibility, which does not converge until 500 iterations in Model (I) and 280 iterations in Model (II). In Fig. 1 (b) and (d), we present the ADMM algorithmic regularity paths obtained by Algorithm 2. In both models, the sparsity levels of the ADMM algorithmic regularity paths increase sharply within only a few iterations as the regularity parameter λ increases, and achieve a relatively smooth transition that explores the range of possible sparse estimates at a fine resolution, and requires substantially less computer time for Algorithm 2 to converge. We also remark here that there are some fluctuations in all four plots of Fig. 1, main because we apply the $(\boldsymbol{B})_+$ operator to ensure the solution to the $\boldsymbol{\Theta}$-subproblem, denoted $\boldsymbol{\Theta}$, to be positive-definite.

In Figs. 2 and 3, we plot the ADMM regularization paths with cool starts in (a), the ADMM regularization paths with warm starts in (b), the ADMM algorithmic regularization paths with a relatively large step size in (c) and a tiny step size in (d), for Models (I) and (II), respectively. In both figures, the ADMM algorithm with warm starts is almost equivalent to the ADMM algorithm with cool starts because the regularization paths exhibit almost identical patterns. The ADMM algorithmic regularization paths with a tiny step size also well approximate the ADMM regularization paths, however, those with a relatively large step-size yields a sequence of sparse models that distinguishes obviously from the sparsity patterns of ADMM algorithm with warm starts. This is mainly because a large change in regularization levels of each step makes that the sparsity levels of the $\boldsymbol{\Theta}$-subproblem after the one-step approximation are not equivalent to that of the solution to (2).

We further compare our algorithms for computing the regularization paths in terms of computational time based on 100 repetitions and each of size $n=50$. The simulations are entirely coded in Matlab and carried out on an AMD 3.30 GHz processor. The results are summarized in Table 1. Notice that, our proposed ADMM algorithmic regularization paths are at least two times faster than the ADMM with warm starts, while the latter is far superior to the state-of-art ADMM with cool starts.

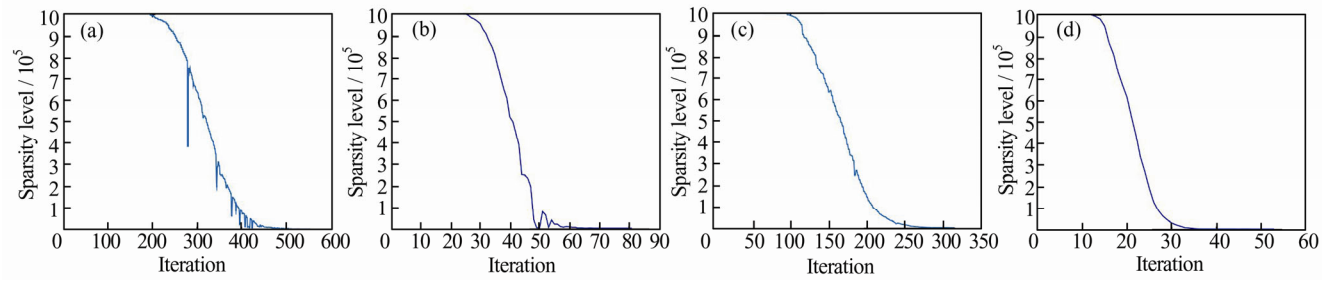

Fig. 1 Sparsity levels measured by the number of nonzero entries of $\hat{\boldsymbol{\Theta}}$

(a) and (c) describe the sparsity levels of the $\boldsymbol{\Theta}$-subproblem iterates of the ADMM algorithm, fit for a fixed λ; (b) and (d) describe the sparsity levels of the $\boldsymbol{\Theta}$-subproblem over the iterates of our proposed ADMM algorithmic regularization paths. (a) and (b) are for Model (I), and (c) and (d) are for Model (II)

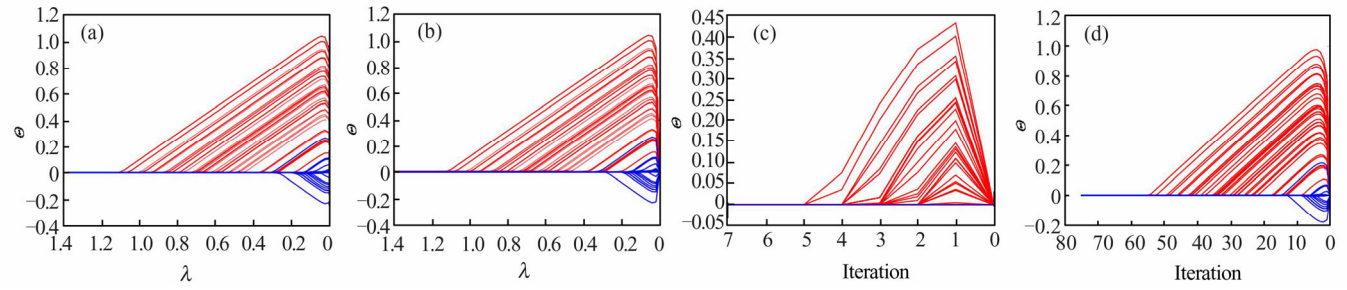

Fig. 2 Simulation results for Model (I)

(a) The ADMM regularization paths with cool starts; (b) The ADMM regularization paths with warm starts; (c) The ADMM algorithmic regularization paths with a relatively large step size; and (d) The ADMM algorithmic regularization paths with a tiny step size. The blue lines denote the false positive variables and the red lines denote true positive variables

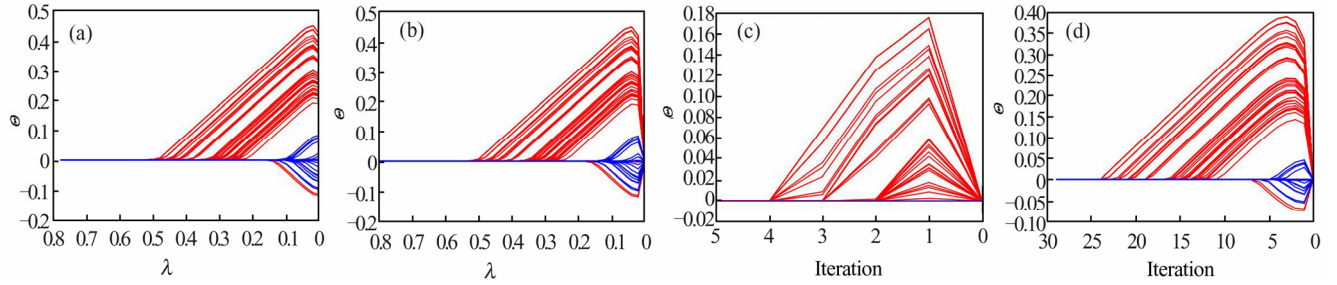

Fig. 3 Simulation results for Model (II)

(a) The ADMM regularization paths with cool starts; (b) The ADMM regularization paths with warm starts; (c) The ADMM algorithmic regularization paths with a relatively large step size; and (d) The ADMM algorithmic regularization paths with a tiny step size. The blue lines denote the false positive variables and the red lines denote true positive variables

Table 1 Timing comparison (averaged over 100 replications) of Algorithm 1 and Algorithm 2 in computing the regularization paths for both models

Algorithm	Dimension p	Time / s	
		Model (I)	Model (II)
1	100	0.557 2	0.309 4
	200	2.428 6	1.413 9
	500	18.870 8	11.161 4
	1 000	643.303 2	473.819 1
2	100	0.227 2	0.124 4
	200	1.035 5	0.601 1
	500	8.466 2	4.786 6
	1 000	316.623 6	177.755 7

3 Conclusion

In this paper, we propose two efficient algorithms to quickly derive the regularization paths for estimating the sparse and large scale positive-definite covariance matrices. Instead of solving the optimization problems over a grid of regularization parameters, which is required by the conventional regularization methods, we propose a one-step approximation to the ADMM algorithm with warm starts to quickly approximate the regularization paths. These new algorithms are easy to implement because no iteration is required. These algorithms both estimate the covariance matrices over a lot of regularization parameters, but ADMM algorithmic regularization paths (Algorithm 2) can quickly outline a sequence of sparse solutions at a fine resolution, and ADMM algorithmic regularization paths are at least two times fast than the ADMM with warm starts.

References

[1] Bickel P J, Levina E. Covariance regularization by thresholding [J]. *The Annals of Statistics*, 2008, **36**(6): 2577-2604.

[2] Cai T T, Zhang C H, Zhou H H, *et al*. Optimal rates of convergence for covariance matrix estimation [J]. *The Annals of Statistics*, 2010, **38**(4): 2118-2144.

[3] Rothman A J, Levina E, Zhu J. Generalized thresholding of large covariance matrices [J]. *Journal of the American Statistical Association*, 2009, **104**(485): 177-186.

[4] Xue L Z, Ma S Q, Zou H. Positive-definite 1-penalized estimation of large covariance matrices [J]. *Journal of the*

American Statistical Association, 2012, **107**(500): 1480-1491.

[5] Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso [J]. *Biostatistics*, 2007, **9**(3): 432-441.

[6] Rothman A J. Positive definite estimators of large covariance matrices [J]. *Biometrika*, 2012, **99**(3): 733-740.

[7] Kang X N, Deng X W. On variable ordination of cholesky-based estimation for a sparse covariance matrix [J]. *The Canadian Journal of Statistics*, 2021, **49**(2): 283-310.

[8] Choi Y G, Lim J, Roy A J P. Fixed support positive-definite modification of covariance matrix estimators via linear shrinkage [J]. *Journal of Multivariate Analysis*, 2019, **171**: 234-249.

[9] Zhang B, Zhou J, Li J. Improved covariance matrix estimators by multi-penalty regularization [C]// 2019 22*th International Conference on Information Fusion* (*FUSION*). Washington D C: IEEE, 2019: 1-7.

[10] Friedman J, Hastie T, Hofling H, *et al*. Pathwise coordinate optimization [J]. *The Annals of Applied Statistics*, 2007, **1**(2): 302-332.

[11] Wu T T, Lange K. Coordinate descent algorithms for lasso penalized regression [J]. *The Annals of Applied Statistics*, 2008, **2**(1): 224-244.

[12] Hu Y, Chi E C, Allen G I. ADMM algorithmic regularization paths for sparse statistical machine learning [C]// *Splitting Methods in Communication*, *Imaging*, *Science*, *and Engineering*. Berlin: Springer-Verlag, 2016: 433-459.

[13] Wahlberg B, Boyd S, Annergren M, *et al*. An ADMM algorithm for a class of total variation regularized estimation problems [J]. *IFAC Proceedings Volumes*, 2012, **45**(16): 83-88.

[14] Mohan K, Chung M, Han S, *et al*. Structured learning of gaussian graphical models [J]. *Advances in Neural Information Processing Systems*, 2012, **2012**:629-637.

[15] Mohan K, London P, Fazel M, *et al*. Node-based learning of multiple Gaussian graphical models [J]. *Journal of Machine Learning Research*, 2014, **15**(1): 445-488.

[16] Chi E C, Lange K. Splitting methods for convex clustering [J]. *Journal of Computational and Graphical Statistics*, 2015, **24**(4): 994-1013.

[17] Boyd S, Parikh N, Chu E, *et al*. Distributed optimization and statistical learning via the alternating direction method of multipliers [J]. *Foundations & Trends in Machine Learning*, 2010, **3**(1): 1-122.

[18] He B S, Yang H, Wang S L. Alternating direction method with self adaptive penalty parameters for monotone variational inequalities [J]. *Journal of Optimization Theory & Applications*, 2000, **106**(2): 337-356.

□