ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095334

LARGE COVARIANCE MATRIX ESTIMATION WITH ORACLE STATISTICAL RATE

Quan Wei and Ziping Zhao

School of Information Science and Technology, ShanghaiTech University, Shanghai, China

ABSTRACT

The ℓ_1 penalized covariance estimator has been widely used for estimating large sparse covariance matrices. It was recognized that ℓ_1 penalty introduces a non-negligible estimation bias, while a proper utilization of non-convex penalty may lead to an estimator with a refined statistical rate of convergence. In this paper, to eliminate the estimation bias we propose to estimate large sparse covariance matrices using the non-convex penalty. It is a challenging task to analyze the theoretical properties of the resulting covariance estimator because popular iterative algorithms for convex optimization no longer have global convergence guarantees for non-convex optimization. To tackle this issue, an efficient algorithm based on the majorization-minimization (MM) is developed by solving a sequence of convex relaxation subproblems. We prove that the proposed estimator computed exactly by the MM-based algorithm achieves the oracle statistical rate under weak assumptions. Our theoretical findings are corroborated through extensive numerical experiments.

***Index Terms*—** Covariance estimation, sparsity, non-convex statistical optimization, majorization-minimization.

1. INTRODUCTION

The estimation of covariance matrices is a fundamental problem in modern multivariate data analysis. It has broad applications in statistics [1], biology [2], finance [3, 4], signal processing [5–7], machine learning [8], etc. When the dimension of the covariance matrix is large, the estimation problem is generally challenging. It is well-known that when the dimension is larger than the sample size, the sample covariance matrix (SCM) is singular, which may cause trouble in real applications. In addition, the number of parameters to be estimated grows quadratically with the dimension of the covariance matrix. Therefore, large covariance matrix estimation has received considerable attention over the past decade.

In order to estimate large covariance matrices effectively, one of the most popular assumptions is sparsity, i.e., a majority of the off-diagonal elements are nearly zero, which reduces the number of parameters to be estimated. The sparsity assumption is reasonable in many real applications [9]. A commonly used method for sparse covariance matrix estimation is thresholding [10–12], which is to set small elements in the SCM to zeros. Although simple, the thresholding covariance estimator is only asymptotically positive definite [10, 11]. In practice, it is more desirable to acquire the positive definiteness in finite samples.

To simultaneously achieve sparsity and positive definiteness, Rothman [13] in his seminal paper suggested to add a log-determinant barrier function into the soft-thresholding (i.e., ℓ_1 penalized) covariance estimation problem [12] and studied an iterative algorithm based on coordinate descent [14]. In [15–17], the authors proposed to obtain a positive definite ℓ_1 penalized covariance estimator by adding a constraint on the smallest eigenvalue of the covariance matrix. In the literature, there are other methods for sparse covariance matrix estimation, e.g., penalized likelihood methods [18,19]. We refer readers to [20, 21] for a comprehensive review.

The ℓ_1 penalized covariance estimator [13, 15–17] has been extensively studied for estimating large sparse covariance matrices in the literature. However, it is now a consensus that the ℓ_1 penalty, e.g., the one used in linear regression, i.e., Lasso [22], introduces a non-negligible estimation bias into the resulting estimator [23]. To alleviate this bias effect, non-convex penalties such as smoothly clipped absolute deviation (SCAD) penalty [23] and minimax concave penalty (MCP) [24] were proposed as alternatives. In [23–25], it has been shown that the non-convex penalized regression is able to eliminate the estimation bias and attain a refined statistical rate of convergence.

Based on these insights, in this paper, we propose to estimate large sparse covariance matrices using the non-convex penalty. It is challenging to analyze the theoretical properties of the resulting covariance estimator due to the non-convexity of the penalty function. If we directly apply the iterative algorithm studied by Rothman [13], then the global optimum may not be obtainable. And for the local optimums, they are in general hard to be characterized. To address the above issue, we develop an efficient multistage convex relaxation algorithm based on majorization-minimization (MM) [26] by solving a sequence of convex subproblems, which can guarantee that an approximate local optimum enjoys the same optimal statistical property of the unobtainable global optimum. We also prove that the proposed estimator computed by the

This work was supported in part by the National Nature Science Foundation of China (NSFC) under Grant 62001295 and in part by the Shanghai Sailing Program under Grant 20YF1430800.

MM-based algorithm achieves the oracle statistical rate under weak assumptions. Numerical experiments verify the theoretical findings and demonstrate the effectiveness of the method.

2. PROBLEM FORMULATION

Given a collection of observations $\left\{\mathbf{x}_i \in \mathbb{R}^d\right\}_{i=1}^n$, the SCM is defined as $\mathbf{S} = \frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$, where $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$ is the sample mean. We consider the following non-convex optimization problem

$$\min_{\mathbf{\Sigma}} \left\{ \frac{1}{2}\left\|\mathbf{\Sigma} - \mathbf{S}\right\|_F^2 - \tau \log\det \mathbf{\Sigma} + \sum_{i \neq j} p_\lambda(|\Sigma_{ij}|) \right\}, \quad (1)$$

where $\tau > 0$ is the barrier parameter fixed at a small value, and $p_\lambda(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$ is a non-convex penalty function with a regularization parameter $\lambda > 0$. The log-determinant barrier function ensures the existence of a positive definite solution [13]. The penalty function $p_\lambda(\cdot)$ is used to encourage a sparse solution. We consider a class of non-convex penalty functions satisfying the following assumptions.

Assumption 1. *The function $p_\lambda(t)$ is defined on $[0, +\infty)$ and satisfies:*

(a) $p_\lambda(t)$ is non-decreasing on $[0, +\infty)$ with $p_\lambda(0) = 0$ and is differentiable almost everywhere on $(0, +\infty)$;

(b) $0 \leq p'_\lambda(t_1) \leq p'_\lambda(t_2) \leq \lambda$ for all $t_1 \geq t_2 \geq 0$ and $\lim_{t\to 0} p'_\lambda(t) = \lambda$.

Prototypical examples of the penalty function $p_\lambda(\cdot)$ in Assumption 1 include SCAD [23] and MCP [24].

3. OPTIMIZATION ALGORITHM

3.1. The MM Algorithmic Framework: A Brief Review

Consider the minimization of a continuous function $F(\mathbf{x})$. Initialized as $\mathbf{x}^{(0)}$, the MM algorithm [26] generates a sequence of feasible points $\{\mathbf{x}^{(k)}\}_{k\geq 1}$ by the following induction. At point $\mathbf{x}^{(k-1)}$, in the majorization step, we design a surrogate function $\bar{F}(\mathbf{x} \mid \mathbf{x}^{(k-1)})$ that locally approximates the objective function $F(\mathbf{x})$, satisfying

$$\begin{cases} \bar{F}(\mathbf{x} \mid \mathbf{x}^{(k-1)}) \geq F(\mathbf{x}), \\ \bar{F}(\mathbf{x}^{(k-1)} \mid \mathbf{x}^{(k-1)}) = F(\mathbf{x}^{(k-1)}). \end{cases}$$

Then, in the minimization step, we update $\mathbf{x}^{(k)}$ as

$$\mathbf{x}^{(k)} \in \arg\min_{\mathbf{x}} \left\{ \bar{F}(\mathbf{x} \mid \mathbf{x}^{(k-1)}) \right\}.$$

3.2. MM-Based Multistage Convex Relaxation Algorithm

We follow the MM framework to solve (1). Here, we find a surrogate function by linearizing $\sum_{i\neq j} p_\lambda(|\Sigma_{ij}|)$. Consequently, we consider a multistage procedure that solves a sequence of convex relaxation subproblems, which is also known as an iteratively reweighted ℓ_1 algorithm. Define $f(\mathbf{\Sigma}) = \frac{1}{2}\left\|\mathbf{\Sigma} - \mathbf{S}\right\|_F^2 - \tau \log\det \mathbf{\Sigma}$. Specifically, starting with an initial value $\widehat{\mathbf{\Sigma}}^{(0)}$, we consider a sequence of convex optimization problems

$$\min_{\mathbf{\Sigma}} \left\{ f(\mathbf{\Sigma}) + \sum_{i\neq j} p'_\lambda(|\widehat{\Sigma}_{ij}^{(k-1)}|)|\Sigma_{ij}| \right\}, \quad 1 \leq k \leq K, \quad (2)$$

where $\widehat{\mathbf{\Sigma}}^{(k)}$ is the optimal solution to the k-th subproblem.

Each subproblem in (2) corresponds to a weighted ℓ_1 penalized covariance estimation problem, which generally can be written into the following form

$$\min_{\mathbf{\Sigma}} \left\{ f(\mathbf{\Sigma}) + \left\|\mathbf{\Lambda} \odot \mathbf{\Sigma}\right\|_{1,\text{off}} \right\}, \quad (3)$$

where $\mathbf{\Lambda}$ is a $d \times d$ matrix of regularization parameters with $\Lambda_{ij} \in [0, \lambda]$. From convex optimization theory, we know that any optimal solution $\widehat{\mathbf{\Sigma}}$ to (3) satisfies the first-order optimality condition $\nabla f(\widehat{\mathbf{\Sigma}}) + \mathbf{\Lambda} \odot \widehat{\mathbf{\Xi}} = \mathbf{0}$ with $\widehat{\mathbf{\Xi}} \in \partial\|\widehat{\mathbf{\Sigma}}\|_{1,\text{off}}$, where $\nabla f(\mathbf{\Sigma}) = \mathbf{\Sigma} - \mathbf{S} - \tau\mathbf{\Sigma}^{-1}$. Since there is no analytical solution to (3), the optimal solution $\widehat{\mathbf{\Sigma}}$ can never be achieved due to numerical optimization error in practice. Instead, we consider to compute an approximate solution to (3).

Definition 2. For a pre-specified tolerance level ε, we say $\widetilde{\mathbf{\Sigma}}$ is an ε-optimal solution to (3) if $\omega_{\mathbf{\Lambda}}(\widetilde{\mathbf{\Sigma}}) \leq \varepsilon$, where

$$\omega_{\mathbf{\Lambda}}(\mathbf{\Sigma}) = \min_{\mathbf{\Xi}\in\partial\|\mathbf{\Sigma}\|_{1,\text{off}}} \left\|\nabla f(\mathbf{\Sigma}) + \mathbf{\Lambda} \odot \mathbf{\Xi}\right\|_{\max}.$$

In view of Definition 2, we use $\widetilde{\mathbf{\Sigma}}^{(k)}$ to denote an ε-optimal solution to the k-th subproblem in (2), given by

$$\min_{\mathbf{\Sigma}} \left\{ f(\mathbf{\Sigma}) + \left\|\mathbf{\Lambda}^{(k-1)} \odot \mathbf{\Sigma}\right\|_{1,\text{off}} \right\}, \quad (4)$$

where $\Lambda_{ij}^{(k-1)} = p'_\lambda(|\widetilde{\Sigma}_{ij}^{(k-1)}|)$. Then, the MM-based multistage convex relaxation algorithm is summarized in Algorithm 1. For simplicity, we start with a trivial initial value $\widetilde{\mathbf{\Sigma}}^{(0)} = \mathbf{I}$. Since $p'_\lambda(|\widetilde{\Sigma}_{ij}^{(0)}|) = p'_\lambda(0) = \lambda$, the first subproblem coincides with the positive definite ℓ_1 penalized covariance estimation problem developed in [13].

To obtain $\widetilde{\mathbf{\Sigma}}^{(k)}$ by solving (4), we apply a proximal gradient method with backtracking line search [27]. We start with[1] $\mathbf{\Sigma}_0^{(k)} = \widetilde{\mathbf{\Sigma}}^{(k-1)}$ and establish the sequence $\{\mathbf{\Sigma}_t^{(k)}\}_{t\geq 1}$ from the proximal gradient method. According to Definition 2, given a prefixed optimization error ε, we stop the proximal gradient algorithm when $\omega_{\mathbf{\Lambda}^{(k-1)}}(\mathbf{\Sigma}_{t+1}^{(k)}) \leq \varepsilon$.

[1] In this paper, the superscript (k) in $\mathbf{\Sigma}_t^{(k)}$ denotes the k-th subproblem in the MM-based algorithm and the subscript t denotes the t-th iteration of the proximal gradient method for solving the subproblems.

Algorithm 1: The MM-Based Multistage Convex Relaxation Algorithm for Solving (1).

```
Input: $\mathbf{S}$, $\tau$, $\lambda$;
Initialize $\widetilde{\mathbf{\Sigma}}^{(0)}$;
for $k = 1, 2, \ldots, K$ do
    $\Lambda_{ij}^{(k-1)} = p'_{\lambda}(|\widetilde{\Sigma}_{ij}^{(k-1)}|)$;
    obtain $\widetilde{\mathbf{\Sigma}}^{(k)}$ by solving problem (4);
    $k = k + 1$;
end
Output: $\widetilde{\mathbf{\Sigma}}^{(K)}$.
```

4. THEORETICAL RESULTS

4.1. Assumptions

We denote the true covariance matrix by $\mathbf{\Sigma}^*$. Let $\mathcal{S}^* = \left\{(i,j) \mid \Sigma_{ij}^* \neq 0\right\}$ be the support set of $\mathbf{\Sigma}^*$ and s^* be its cardinality, i.e., $|\mathcal{S}^*| = s^*$. Recall that $p_\lambda(t)$ is a non-convex penalty function and $p'_\lambda(t)$ is its derivative. In the following, we impose some mild conditions on the true covariance matrix $\mathbf{\Sigma}^*$ and the function $p'_\lambda(t)$.

Assumption 3. *For the true covariance matrix $\mathbf{\Sigma}^*$, there exists $\kappa \geq 1$ such that $0 < \frac{1}{\kappa} \leq \lambda_{\min}(\mathbf{\Sigma}^*) \leq \lambda_{\max}(\mathbf{\Sigma}^*) \leq \kappa < \infty$.*

The bounded eigenvalue condition on the true covariance matrix in Assumption 3 is standard in the existing literature on sparse covariance matrix estimation problems [13, 16, 18].

Assumption 4. *The function $p'_\lambda(t)$ satisfies:*

(a) There exists an $\alpha > 0$ such that $p'_\lambda(t) = 0$ for $t \geq \alpha\lambda$;

(b) There exists some $c \in (0, \alpha)$ such that $p'_\lambda(c\lambda) \geq \frac{\lambda}{2}$.

In Assumption 4, the first condition holds for various non-convex penalty functions including the popular choices: SCAD [23] and MCP [24]; the second condition can always hold due to $p'_\lambda(0) = \lambda$ and $p'_\lambda(\alpha\lambda) = 0$.

Assumption 5. *The true covariance matrix $\mathbf{\Sigma}^*$ satisfies $\|\mathbf{\Sigma}_{\mathcal{S}}^*\|_{\min} = \min_{(i,j)\in\mathcal{S}^*} |\Sigma_{ij}^*| \geq (\alpha + c)\lambda \gtrsim \lambda$, where α and c are the same to those given in Assumption 4.*

Assumption 5 is referred to as the *minimum signal strength condition*, which has been widely employed in the analysis of non-convex penalized regression problems [23–25]. This condition is very mild because in our statistical analysis, the tuning parameter λ will be shown to be in the order of $\sqrt{\frac{\log d}{n}}$ that could be very small when the sample size n is large.

4.2. Statistical Analysis

We now present the main theorem, which demonstrates the contraction property of the solution path $\{\widetilde{\mathbf{\Sigma}}^{(k)}\}_{k\geq 1}$.

Theorem 6. *Suppose that Assumptions 1, 3, 4, and 5 hold. If $\lambda \geq 2\left(\|\nabla f(\mathbf{\Sigma}^*)\|_{\max} + \varepsilon\right)$, then the ε-optimal solution $\widetilde{\mathbf{\Sigma}}^{(k)}$ $(1 \leq k \leq K)$ satisfies the following contraction property:*

$$\left\|\widetilde{\mathbf{\Sigma}}^{(k)} - \mathbf{\Sigma}^*\right\|_F \leq \underbrace{\|(\nabla f(\mathbf{\Sigma}^*))_{\mathcal{S}^*}\|_F}_{\textit{oracle rate}} + \underbrace{\varepsilon\sqrt{s^*}}_{\textit{optimization error}} + \delta \underbrace{\left\|\widetilde{\mathbf{\Sigma}}^{(k-1)} - \mathbf{\Sigma}^*\right\|_F}_{\textit{contraction}},$$

where $\delta \in (0,1)$ is the contraction parameter.

Remark 7. The oracle rate refers to the statistical convergence rate of the oracle estimator. Based on known $\mathcal{S}^*$, the oracle estimator is defined as $\widehat{\mathbf{\Sigma}}^O = \arg\min_{\mathbf{\Sigma}:\, \mathbf{\Sigma}_{\overline{\mathcal{S}^*}}=\mathbf{0}} f(\mathbf{\Sigma})$. Then, we obtain the rate for $\widehat{\mathbf{\Sigma}}^O$ as $\left\|\widehat{\mathbf{\Sigma}}^O - \mathbf{\Sigma}^*\right\|_F \lesssim \|(\nabla f(\mathbf{\Sigma}^*))_{\mathcal{S}^*}\|_F$.

Theorem 6 demonstrates that the estimation error between the ε-optimal solution $\widetilde{\mathbf{\Sigma}}^{(k)}$ and the true value $\mathbf{\Sigma}^*$ is bounded by three terms, namely, the oracle rate, the optimization error, and a contraction term. Because the contraction property is induced by the MM-based multistage convex relaxation algorithm, in order to achieve the oracle rate, we shall carefully choose the optimization error such that $\varepsilon \leq \frac{\|(\nabla f(\mathbf{\Sigma}^*))_{\mathcal{S}^*}\|_F}{\sqrt{s^*}}$ and make k large enough. Next, we further give the explicit statistical rate of convergence under the sub-Gaussian design.

Corollary 8. *Let $\mathbf{x}$ be a sub-Gaussian random vector with zero mean and covariance $\mathbf{\Sigma}^*$ and $\{\mathbf{x}_i\}_{i=1}^n$ be a collection of independent and identically distributed samples from $\mathbf{x}$. Suppose that Assumptions 1, 3, 4, and 5 hold. If $\lambda \asymp \sqrt{\frac{\log d}{n}}$, $\tau \lesssim \sqrt{\frac{s^*}{n}}\left\|(\mathbf{\Sigma}^*)^{-1}\right\|_F^{-1}$, $\varepsilon \lesssim \sqrt{\frac{1}{n}}$, and $K \gtrsim \log(\lambda\sqrt{n}) \gtrsim \log\log d$, then the ε-optimal solution $\widetilde{\mathbf{\Sigma}}^{(K)}$ satisfies $\left\|\widetilde{\mathbf{\Sigma}}^{(K)} - \mathbf{\Sigma}^*\right\|_F \lesssim \sqrt{\frac{s^*}{n}}$ with high probability.*

Corollary 8 is a direct consequence of Theorem 6, which implies that under weak assumptions, we just need to solve no more than approximately $\log\log d$ convex problems, i.e., the problem (4), to achieve the oracle rate $\sqrt{\frac{s^*}{n}}$. In the literature, the sparse covariance matrix estimation problem has been studied via ℓ_1 penalty, with the order of $\sqrt{\frac{s^* \log d}{n}}$ [13, 15, 16]. It is easy to see that the proposed estimator achieves a faster statistical rate of convergence (matching the oracle rate) than the existing ones with ℓ_1 penalty in the Frobenius norm.

5. NUMERICAL EXPERIMENTS

We compare our proposed large covariance matrix estimator with several existing ones [12, 13, 15]. More specifically, we consider the following four methods in the simulation.

- `TCE-L1` [12]: thresholding covariance estimator (TCE) with ℓ_1 penalty (a.k.a. soft-TCE).

Table 1. Quantitative comparison among four different methods for banded, block, and Toeplitz settings.

Covariance Model	Metrics	n	`TCE-L1`	`TCE-MCP`	`PDSCE-L1`	`PDSCE-MCP` (prop.)
Banded	$\|\cdot\|_F$	50	11.2293 (0.0724)	10.5731 (0.0939)	11.1628 (0.0687)	10.5391 (0.0725)
		100	7.2799 (0.0528)	6.4581 (0.0530)	7.2534 (0.0441)	6.4467 (0.0429)
		150	6.0543 (0.0494)	5.1819 (0.0492)	6.0273 (0.0415)	5.1765 (0.0392)
	P. D.	50	8/100	0/100	100/100	100/100
		100	13/100	0/100	100/100	100/100
		150	24/100	0/100	100/100	100/100
Block	$\|\cdot\|_F$	50	9.7928 (0.0537)	8.7781 (0.0539)	9.7371 (0.0809)	8.7338 (0.0649)
		100	6.1063 (0.0424)	4.2924 (0.0496)	6.0934 (0.0412)	4.2698 (0.0423)
		150	5.1459 (0.0388)	3.3947 (0.0411)	5.1178 (0.0379)	3.3719(0.0382)
	P. D.	50	89/100	0/100	100/100	100/100
		100	100/100	66/100	100/100	100/100
		150	100/100	100/100	100/100	100/100
Toeplitz	$\|\cdot\|_F$	50	10.3239 (0.0396)	9.4798 (0.0410)	10.2834 (0.0396)	9.4225 (0.0337)
		100	6.6634 (0.0329)	6.4641 (0.0323)	6.6474 (0.0329)	6.4234 (0.0257)
		150	5.8295 (0.0294)	5.5686 (0.0229)	5.8055 (0.0285)	5.5525 (0.0215)
	P. D.	50	97/100	0/100	100/100	100/100
		100	100/100	0/100	100/100	100/100
		150	100/100	0/100	100/100	100/100

- `TCE-MCP` [12]: TCE with MCP.
- `PDSCE-L1` [13, 15]: positive definite sparse covariance estimator (PDSCE) with ℓ_1 penalty.
- `PDSCE-MCP` (proposed): PDSCE with MCP.[2]

We generate $n = \{50, 100, 150\}$ independent data points from a d-dimensional Gaussian random variable with zero mean and population covariance $\mathbf{\Sigma}^*$ for $d = 100$. Following [13, 15–17], we consider three typical covariance models.

- Banded matrix: $\Sigma^*_{ij} = \begin{cases} 1 - \frac{|i-j|}{10} & |i-j| \leq 10, \\ 0 & \text{otherwise.} \end{cases}$
- Block matrix: The indices $1, 2, \ldots, d$ are partitioned into 10 ordered groups of equal size with

$$\Sigma^*_{ij} = \begin{cases} 1 & i = j, \\ 0.6 & i \text{ and } j\ (i \neq j) \text{ belong to the same group,} \\ 0 & \text{otherwise.} \end{cases}$$

- Toeplitz matrix: $\Sigma^*_{ij} = 0.75^{|i-j|}$.

For all four methods, the tuning parameter λ is chosen by five-fold cross-validation. For two PDSCEs, the barrier parameter τ is set to be 10^{-4} as suggested in [13]. The resulting covariance estimates are compared with the population covariance matrix $\mathbf{\Sigma}^*$ using the Frobenius norm as evaluation metrics. All results are averaged on 100 Monte Carlo realizations. We report the averaged estimation errors in Table 1 with the corresponding standard errors in parentheses.

From the simulation results, we can see that for all four methods, the estimators with MCP uniformly achieve better estimation performance than the counterparts with ℓ_1 penalty in terms of Frobenius norm error. This supports our theoretical results that non-convex penalty can reduce the covariance estimation error. Among those estimators with the same kind of penalty, TCE and PDSCE achieve similar estimation performance and in particularly PDSCE slightly outperforms TCE. Furthermore, PDSCE always guarantees the positive definiteness of the estimated covariance matrices and in contrast, TCE almost never delivers positive definite estimates, especially for the banded covariance matrix, which is also reported in Table 1. Such results are similar to those obtained by [13, 15–17]. Among all four methods, the proposed estimator achieves the best estimation performance.

6. CONCLUSION

In this paper, we have proposed a novel approach for large sparse covariance matrix estimation using the non-convex penalty and presented both theoretical and empirical results. To the best of our knowledge, this is the first work to improve the statistical rate of convergence under the Frobenius norm for large sparse covariance matrix estimation. It is also interesting to incorporate the iteration complexity analysis of the proximal gradient method for solving the subproblems in the MM-based algorithm into the property characterization of the proposed estimator. We leave this for future work.

[2] The function $p_\lambda(\cdot)$ for MCP is defined through its derivative: $p'_\lambda(t) = \max\left(\lambda - \frac{t}{a}, 0\right)$, where $a > 1$ [24]. We set $a = 2$ in all experiments.

7. REFERENCES

[1] A. P. Dempster, “Covariance selection,” *Biometrics*, vol. 28, no. 1, pp. 157–175, 1972.

[2] J. Schäfer and K. Strimmer, “A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics,” *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1, 2005.

[3] Z. Zhao and D. P. Palomar, “Mean-reverting portfolio with budget constraint,” *IEEE Transactions on Signal Processing*, vol. 66, no. 9, pp. 2342–2357, 2018.

[4] Z. Zhao, R. Zhou, and D. P. Palomar, “Optimal mean-reverting portfolio with leverage constraint for statistical arbitrage in finance,” *IEEE Transactions on Signal Processing*, vol. 67, no. 7, pp. 1681–1695, 2019.

[5] E. Ollila, D. E. Tyler, V. Koivunen, and H. V. Poor, “Complex elliptically symmetric distributions: Survey, new results and applications,” *IEEE Transactions on Signal Processing*, vol. 60, no. 11, pp. 5597–5625, 2012.

[6] Y. Sun, P. Babu, and D. P. Palomar, “Robust estimation of structured covariance matrix for heavy-tailed elliptical distributions,” *IEEE Transactions on Signal Processing*, vol. 64, no. 14, pp. 3576–3590, 2016.

[7] E. Ollila, D. P. Palomar, and F. Pascal, “Shrinking the eigenvalues of m-estimators of covariance matrix,” *IEEE Transactions on Signal Processing*, vol. 69, pp. 256–269, 2021.

[8] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[9] W. B. Wu and M. Pourahmadi, “Nonparametric estimation of large covariance matrices of longitudinal data,” *Biometrika*, vol. 90, no. 4, pp. 831–844, 2003.

[10] P. J. Bickel and E. Levina, “Covariance regularization by thresholding,” *The Annals of Statistics*, vol. 36, no. 6, pp. 2577–2604, 2008.

[11] N. El Karoui, “Operator norm consistent estimation of large-dimensional sparse covariance matrices,” *The Annals of Statistics*, vol. 36, no. 6, pp. 2717–2756, 2008.

[12] A. J. Rothman, E. Levina, and J. Zhu, “Generalized thresholding of large covariance matrices,” *Journal of the American Statistical Association*, vol. 104, no. 485, pp. 177–186, 2009.

[13] A. J. Rothman, “Positive definite estimators of large covariance matrices,” *Biometrika*, vol. 99, no. 3, pp. 733–740, 2012.

[14] S. J. Wright, “Coordinate descent algorithms,” *Mathematical Programming*, vol. 151, no. 1, pp. 3–34, 2015.

[15] L. Xue, S. Ma, and H. Zou, “Positive-definite ℓ_1-penalized estimation of large covariance matrices,” *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1480–1491, 2012.

[16] H. Liu, L. Wang, and T. Zhao, “Sparse covariance matrix estimation with eigenvalue constraints,” *Journal of Computational and Graphical Statistics*, vol. 23, no. 2, pp. 439–459, 2014.

[17] Y. Cui, C. Leng, and D. Sun, “Sparse estimation of high-dimensional correlation matrices,” *Computational Statistics & Data Analysis*, vol. 93, pp. 390–403, 2016.

[18] C. Lam and J. Fan, “Sparsistency and rates of convergence in large covariance matrix estimation,” *The Annals of Statistics*, vol. 37, no. 6B, pp. 4254–4278, 2009.

[19] J. Bien and R. J. Tibshirani, “Sparse estimation of a covariance matrix,” *Biometrika*, vol. 98, no. 4, pp. 807–820, 2011.

[20] M. Pourahmadi, *High-Dimensional Covariance Estimation*. John Wiley & Sons, 2013.

[21] J. Fan, Y. Liao, and H. Liu, “An overview of the estimation of large covariance and precision matrices,” *The Econometrics Journal*, vol. 19, no. 1, pp. C1–C32, 2016.

[22] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[23] J. Fan and R. Li, “Variable selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[24] C.-H. Zhang, “Nearly unbiased variable selection under minimax concave penalty,” *The Annals of Statistics*, vol. 38, no. 2, pp. 894–942, 2010.

[25] J. Fan, H. Liu, Q. Sun, and T. Zhang, “I-lamm for sparse learning: Simultaneous control of algorithmic complexity and statistical error,” *The Annals of Statistics*, vol. 46, no. 2, pp. 814–841, 2018.

[26] Y. Sun, P. Babu, and D. P. Palomar, “Majorization-minimization algorithms in signal processing, communications, and machine learning,” *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 794–816, 2017.

[27] Y. Nesterov, “Gradient methods for minimizing composite functions,” *Mathematical Programming*, vol. 140, no. 1, pp. 125–161, 2013.