Proceedings of IEEE Statistical Signal Processing Workshop, Ann Arbor, MI. Aug 5-7, 2012.

KRONECKER GRAPHICAL LASSO

Theodoros Tsiligkaridis[*], *Alfred O. Hero III*[*,†] *and Shuheng Zhou*[†,*]

University of Michigan, [*]EECS Dept. and [†] Dept. Statistics, Ann Arbor, USA
{ttsili,hero,shuhengz}@umich.edu

ABSTRACT

We consider high-dimensional estimation of a (possibly sparse) Kronecker-decomposable covariance matrix given i.i.d. Gaussian samples. We propose a sparse covariance estimation algorithm, Kronecker Graphical Lasso (KGlasso), for the high dimensional setting that takes advantage of structure and sparsity. Convergence and limit point characterization of this iterative algorithm is established. Compared to standard Glasso, KGlasso has low computational complexity as the dimension of the covariance matrix increases. We derive a tight MSE convergence rate for KGlasso and show it strictly outperforms standard Glasso and FF. Simulations validate these results and shows that KGlasso outperforms the maximum-likelihood solution (FF), in the high-dimensional small-sample regime.

Index Terms— sparsity, structured covariance estimation, penalized maximum likelihood, graphical lasso

1. INTRODUCTION

Covariance estimation is a problem of great interest in many different disciplines, including machine learning, signal processing, economics and bioinformatics. In this paper we consider covariance estimation in the multivariate Gaussian model under the separable positive definite $pf \times pf$ covariance matrix assumption:

$$\mathbf{\Sigma}_0 = \mathbf{A}_0 \otimes \mathbf{B}_0 \tag{1}$$

where $\mathbf{A}_0$ is a $p \times p$ positive definite matrix and $\mathbf{B}_0$ is an $f \times f$ positive definite matrix. Model (1) arises in channel modeling for MIMO wireless communications, where $\mathbf{A}_0$ is a transmit covariance matrix and $\mathbf{B}_0$ is a receive covariance matrix, and in other applications, see [1]. Let $\mathbf{\Theta}_0 := \mathbf{\Sigma}_0^{-1}$ denote the inverse covariance, or precision matrix. As compared to the standard saturated (unstructured) model, the number of independent parameters in (1) is reduced from $\Theta(p^2 f^2)$ to $\Theta(p^2) + \Theta(f^2)$. Furthermore, as shown [1], factorization (1) results in a significant reduction in estimation mean squared error and in estimator computational complexity. In this paper we propose estimation of a sparse version of the Kronecker product model (1) resulting in even more significant performance improvements than for the saturated model studied in [1].

Under model (1), the joint probability distribution of the measurements can be represented by an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the vertex set (each vertex corresponding to a variable) and $\mathcal{E}$ is the edge set. If, given all the other variables, the ith variable is conditionally independent of the jth variable, then $(i, j) \notin \mathcal{E}$ [2]. Estimating an undirected Gaussian graphical model is equivalent to estimating the inverse covariance matrix. Penalized likelihood estimators for Gaussian graphical models, such as the graphical lasso (Glasso) have been proposed [3, 4, 5]. The maximum-likelihood (ML) estimator of the Kronecker product (1) has been studied in [1, 6]. While the ML estimator has no known closed-form solution, an approximation to the solution can be iteratively computed via an alternating algorithm: the flip-flop (FF) algorithm [1, 6].

To our knowledge, ML estimation for the situation where the Kronecker component matrices are themselves sparse has not been studied. In addition to the Kronecker factorization, we exploit sparsity in order to derive better estimators, especially for the large-dimension small-sample regime. In this paper, we propose an ℓ_1-penalized likelihood estimator for the sparse Kronecker product case.

Statistical consistency is guaranteed, i.e., the estimator converges in probability to the true inverse covariance matrix $\mathbf{\Theta}_0$ asymptotically as the number of samples and dimensions of Kronecker factor matrices grows to infinity. The main contribution is the derivation of the high-dimensional MSE convergence rates for KGlasso. When both Kronecker factors are sparse, it is shown that KGlasso *strictly* outperforms FF and naive Glasso in MSE, and the performance improvement can be very significant. Simulations show that KGlasso exhibits superior empirical performance.

2. NOTATION

For a square matrix $\mathbf{M}$, define $|\mathbf{M}|_1 = \|\text{vec}(\mathbf{M})\|_1$ and $|\mathbf{M}|_\infty = \|\text{vec}(\mathbf{M})\|_\infty$, where $\text{vec}(\mathbf{M})$ denotes the vectorized form of $\mathbf{M}$ (concatenation of columns into a vector). $\|\mathbf{M}\|_2$ is the spectral norm of $\mathbf{M}$. $\mathbf{M}_{i,j}$ and $[\mathbf{M}]_{i,j}$ are the (i, j)th el-

The research reported here was partially supported by a grant from ARO, W911NF-11-1-0391, and a Rackham Merit fellowship to the first author.

ement of $\mathbf{M}$. Let the inverse transformation (from a vector to a matrix) be defined as: $\text{vec}^{-1}(\mathbf{x}) = \mathbf{X}$, where $\mathbf{x} = \text{vec}(\mathbf{X})$. Define the $pf \times pf$ permutation operator $\mathbf{K}_{p,f}$ such that $\mathbf{K}_{p,f}\text{vec}(\mathbf{N}) = \text{vec}(\mathbf{N}^T)$ for any $p \times f$ matrix $\mathbf{N}$. For a symmetric matrix $\mathbf{M}$, $\lambda(\mathbf{M})$ will denote the vector of real eigenvalues of $\mathbf{M}$ and define $\lambda_{max}(\mathbf{M}) = \|\mathbf{M}\|_2 = \max \lambda_i(\mathbf{M})$ for p.d. symmetric matrix, and $\lambda_{min}(\mathbf{M}) = \min \lambda_i(\mathbf{M})$.

For a matrix $\mathbf{M}$ of size $pf \times pf$, let $\{\mathbf{M}(i,j)\}_{i,j=1}^{p}$ denote its $f \times f$ block submatrices, where each block submatrix is $\mathbf{M}(i,j) = [\mathbf{M}]_{(i-1)f+1:if,(j-1)f+1:jf}$. Also let $\{\overline{\mathbf{M}}(k,l)\}_{k,l=1}^{f}$ denote the $p \times p$ block submatrices of the permuted matrix $\overline{\mathbf{M}} = \mathbf{K}_{p,f}^T\mathbf{M}\mathbf{K}_{p,f}$.

Define the set of symmetric matrices S^p and the set of symmetric positive definite (pd) matrices S_{++}^p.

3. GRAPHICAL LASSO FRAMEWORK

Available are n i.i.d. multivariate Gaussian observations $\{\mathbf{z}_t\}_{t=1}^n$, where $\mathbf{z}_t \in \mathbb{R}^{pf}$ has zero-mean and covariance equal to $\mathbf{\Sigma} = \mathbf{A}_0 \otimes \mathbf{B}_0$. The log-likelihood function is proportional to:

$$l(\mathbf{\Sigma}) := \log\det(\mathbf{\Sigma}^{-1}) - \text{tr}(\mathbf{\Sigma}^{-1}\hat{\mathbf{S}}_n), \tag{2}$$

where $\mathbf{\Sigma}$ is the positive definite covariance matrix and $\hat{\mathbf{S}}_n = \frac{1}{n}\sum_{t=1}^n \mathbf{z}_t\mathbf{z}_t^T$ is the sample covariance matrix. Recent work [7, 8] has considered ℓ_1-penalized maximum likelihood estimators for the saturated model where $\mathbf{\Sigma}$ belongs to the unrestricted cone of positive definite matrices. These estimators are known as graphical lasso (Glasso) estimators and are the solution to the ℓ_1-penalized minimization problem:

$$\hat{\mathbf{\Sigma}}_n \in \arg\min_{\mathbf{\Sigma} \in S_{++}^p} \{-l(\mathbf{\Sigma}) + \lambda|\mathbf{\Sigma}^{-1}|_1\}, \tag{3}$$

where $\lambda \geq 0$ is a regularization parameter. If $\lambda > 0$ and $\hat{\mathbf{S}}_n$ is positive definite, then $\hat{\mathbf{\Sigma}}_n$ in (3) is the unique minimizer.

A fast iterative algorithm, based on a block coordinate descent approach, exhibiting a computational complexity $\mathcal{O}((pf)^3)$, was developed in [8] to solve the convex program (3). Under the assumption $\lambda \asymp \sqrt{\frac{\log(pf)}{n}}$ solution of (3) was shown to have high dimensional convergence rate [3]:

$$\|\mathbf{G}(\hat{\mathbf{S}}_n, \lambda) - \mathbf{\Theta}_0\|_F = O_P\left(\sqrt{\frac{(pf+s)\log(pf)}{n}}\right) \tag{4}$$

where s is an upper bound on the number of non-zero off-diagonal elements of $\mathbf{\Theta}_0$. When $s = O(pf)$, this rate is better than the non-regularized sample covariance estimator:

$$\|\hat{\mathbf{S}}_n - \mathbf{\Sigma}_0\|_F = O_P\left(\sqrt{\frac{p^2f^2}{n}}\right). \tag{5}$$

4. KRONECKER GRAPHICAL LASSO

Let $\mathbf{\Sigma}_0 := \mathbf{A}_0 \otimes \mathbf{B}_0$ denote the true covariance matrix, where $\mathbf{A}_0 := \mathbf{X}_0^{-1}$ and $\mathbf{B}_0 = \mathbf{Y}_0^{-1}$ are the true Kronecker factors. Let $\mathbf{A}_{init}$ denote the initial guess of $\mathbf{A}_0 = \mathbf{X}_0^{-1}$.

Define $J(\mathbf{X}, \mathbf{Y})$ as the negative log-likelihood

$$\begin{aligned} J(\mathbf{X}, \mathbf{Y}) = \text{tr}((\mathbf{X} \otimes \mathbf{Y})\hat{\mathbf{S}}_n) - f\log\det(\mathbf{X}) \\ - p\log\det(\mathbf{Y}) \end{aligned} \tag{6}$$

Although the objective (6) is not jointly convex in $(\mathbf{X}, \mathbf{Y})$, it is biconvex. This motivates the flip-flop algorithm [1]. Adapting the notation from [1], define the mappings $\hat{\mathbf{A}}(\cdot), \hat{\mathbf{B}}(\cdot)$:

$$\underbrace{\hat{\mathbf{A}}(\mathbf{B})}_{p \times p} = \frac{1}{f}\sum_{k,l=1}^{f} [\mathbf{B}^{-1}]_{k,l}\overline{\hat{\mathbf{S}}_n}(l,k), \tag{7}$$

$$\underbrace{\hat{\mathbf{B}}(\mathbf{A})}_{f \times f} = \frac{1}{p}\sum_{i,j=1}^{p} [\mathbf{A}^{-1}]_{i,j}\hat{\mathbf{S}}_n(j,i), \tag{8}$$

where $\overline{\hat{\mathbf{S}}}_n = \mathbf{K}_{p,f}^T\hat{\mathbf{S}}_n\mathbf{K}_{p,f}$. For fixed $\mathbf{B} \in S_{++}^f$, $\hat{\mathbf{A}}(\mathbf{B})$ in (7) is the minimizer of $J(\mathbf{A}^{-1}, \mathbf{B}^{-1})$ over $\mathbf{A} \in S_{++}^p$. A similar interpretation holds for (8). The flip-flop algorithm [1] starts with some arbitrary p.d. matrix $\mathbf{A}_{init}$ and computes $\mathbf{B}$ using (8), then $\mathbf{A}$ using (7), and repeats until convergence. The FF algorithm does not account for sparsity.

If $\mathbf{\Theta}_0 = \mathbf{X}_0 \otimes \mathbf{Y}_0$ is a sparse matrix, which implies that at least one of $\mathbf{X}_0$ or $\mathbf{Y}_0$ is sparse, one can penalize the outputs of the flip-flop algorithm and iteratively minimize

$$J_\lambda(\mathbf{X}, \mathbf{Y}) = J(\mathbf{X}, \mathbf{Y}) + \bar{\lambda}_X|\mathbf{X}|_1 + \bar{\lambda}_Y|\mathbf{Y}|_1. \tag{9}$$

This leads to an algorithm that we call KGlasso (see Algorithm 1), which sparsifies the Kronecker factors in proportion to the parameters $\bar{\lambda}_X, \bar{\lambda}_Y > 0$. This additive penalty was first proposed in [9] in the context of missing data.

Algorithm 1 Kronecker Graphical Lasso (KGlasso)

```
Input: $\hat{\mathbf{S}}_n$, $p$, $f$, $n$, $\bar{\lambda}_X > 0$, $\bar{\lambda}_Y > 0$
Output: $\hat{\mathbf{\Theta}}_{KGlasso}$
Initialize $\mathbf{A}_{init}$ to be positive definite satisfying Assumption 1.
$\check{\mathbf{X}} \leftarrow \mathbf{A}_{init}^{-1}$
repeat
    $\hat{\mathbf{B}} \leftarrow \frac{1}{p}\sum_{i,j=1}^{p}[\check{\mathbf{X}}]_{i,j}\hat{\mathbf{S}}_n(j,i)$ (see Eq. (7))
    $\check{\mathbf{Y}} \leftarrow \mathbf{G}(\hat{\mathbf{B}}, \frac{\bar{\lambda}_Y}{p})$, where $\mathbf{G}(\cdot,\cdot)$ is defined in (10)
    $\hat{\mathbf{A}} \leftarrow \frac{1}{f}\sum_{k,l=1}^{f}[\check{\mathbf{Y}}]_{k,l}\overline{\hat{\mathbf{S}}_n}(l,k)$ (see Eq. (8))
    $\check{\mathbf{X}} \leftarrow \mathbf{G}(\hat{\mathbf{A}}, \frac{\bar{\lambda}_X}{f})$
until convergence
$\hat{\mathbf{\Theta}}_{KGlasso} \leftarrow \check{\mathbf{X}} \otimes \check{\mathbf{Y}}$
```

The Glasso mapping (3) is written as $\mathbf{G}(\cdot, \lambda) : S^d \to S^d$,

$$\mathbf{G}(\mathbf{T}, \lambda) = \arg \min_{\mathbf{\Theta} \in S^d_{++}} \left\{ \mathrm{tr}(\mathbf{\Theta}\mathbf{T}) - \log\det(\mathbf{\Theta}) + \lambda|\mathbf{\Theta}|_1 \right\}. \tag{10}$$

As compared to the $\mathcal{O}(p^4 f^4)$ computational complexity of Glasso [8], KGlasso has a computational complexity of only $\mathcal{O}(p^4 + f^4)$.

Assuming $\hat{\mathbf{S}}_n$ is p.d., KGlasso converges to a critical point of the objective function [10]. Under a mild assumption on the starting point, KGlasso can be shown to converge to a local minimum [10].

5. HIGH DIMENSIONAL CONSISTENCY OF FF

In this section, we show that the flip-flop (FF) algorithm achieves the optimal (non-sparse) statistical convergence rate of $O_P\left(\sqrt{\frac{p^2+f^2}{n}}\right)$ (up to a log-factor). This result (Thm. 1) allows us to establish that the proposed KGlasso has significantly improved MSE convergence rate (Thm 2). We make the following standard assumption on the spectra of the Kronecker factors.

Assumption 1. *Uniformly Bounded Spectra*
There exist absolute constants $\underline{k}_A, \overline{k}_A, \underline{k}_B, \overline{k}_B, \underline{k}_{A_{init}}, \overline{k}_{A_{init}}$ such that:

1a. $0 < \underline{k}_A \leq \lambda_{min}(\mathbf{A}_0) \leq \lambda_{max}(\mathbf{A}_0) \leq \overline{k}_A < \infty$
1b. $0 < \underline{k}_B \leq \lambda_{min}(\mathbf{B}_0) \leq \lambda_{max}(\mathbf{B}_0) \leq \overline{k}_B < \infty$
2. $0 < \underline{k}_{A_{init}} \leq \lambda_{min}(\mathbf{A}_{init}) \leq \lambda_{max}(\mathbf{A}_{init}) \leq \overline{k}_{A_{init}} < \infty$

Let $\mathbf{R}_{FF}(3) := \hat{\mathbf{A}}(\hat{\mathbf{B}}(\mathbf{A}_{init})) \otimes \hat{\mathbf{B}}(\hat{\mathbf{A}}(\hat{\mathbf{B}}(\mathbf{A}_{init})))$ denote the 3-step (noniterative) version of the flip-flop algorithm [1]. More generally, let $\hat{\mathbf{R}}_{FF}(k)$ denote the k-step version of the flip-flop algorithm. Let $\mathbf{\Theta}_{FF}(k) = (\mathbf{R}_{FF}(k))^{-1}$.

Theorem 1. *Let $\mathbf{A}_0, \mathbf{B}_0$, and $\mathbf{A}_{init}$ satisfy Assumption 1 and define $M = \max(p, f, n)$. Assume $p \geq f \geq 2$ and $p \log M \leq C''n$ for some finite constant $C'' > 0$. Finally, assume $n \geq \frac{p}{f} + 1$. Then, for $k \geq 2$ finite,*

$$\|\mathbf{\Theta}_{FF}(k) - \mathbf{\Theta}_0\|_F = O_P\left(\sqrt{\frac{(p^2+f^2)\log M}{n}}\right) \tag{11}$$

as $n \to \infty$.

Proof. Due to space limitations the proof is given in [10]. □

The bound (11) specifies the rate of reduction of the estimation error for the multi-iteration FF algorithm, which includes the three step FF algorithm ($k = 3$) [1] as a special case. The error reduction decreases as long as p and f do not increase too quickly in n.

Note that (11) specifies a faster rate than that of the naive sample covariance matrix estimator (5). Furthemore, since the computational complexity for FF is $\mathcal{O}(p^2 + f^2)$ which is less than the $\mathcal{O}(p^2 f^2)$ complexity of SCM, by exploiting Kronecker structure FF simultaneously achieves improved MSE performance and reduced computational complexity.

6. HIGH DIMENSIONAL CONSISTENCY OF KGLASSO

In this section, high dimensional consistency is established for KGlasso as $n, p, f \to \infty$.

Define $\mathbf{\Theta}_{KGlasso}(k)$ as the output of the kth KGlasso iteration.

Theorem 2. *Let $\mathbf{A}_0, \mathbf{B}_0, \mathbf{A}_{init}$ satisfy Assumption 1. Let $M = \max(p, f, n)$. Let $\bar{\lambda}_Y^{(1)} \asymp p\sqrt{\frac{\log M}{np}}$ and $\bar{\lambda}_X^{(k)} \asymp \left(\frac{1}{\sqrt{p}} + \frac{1}{\sqrt{f}}\right) f\sqrt{\frac{\log M}{n}}, \bar{\lambda}_Y^{(k')} \asymp \left(\frac{1}{\sqrt{p}} + \frac{1}{\sqrt{f}}\right) p\sqrt{\frac{\log M}{n}}$ as $p, f, n \to \infty$ for all $k \geq 1$ and $k' \geq 2$. Assume sparse $\mathbf{X}_0$ and $\mathbf{Y}_0$, i.e. $s_{X_0} = O(p), s_{Y_0} = O(f)$. Assume $\max\left(\frac{p}{f}, \frac{f}{p}\right) \log M = o(n)$. Then, for $k \geq 2$ finite, we have*

$$\|\mathbf{\Theta}_{KGlasso}(k) - \mathbf{\Theta}_0\|_F = O_P\left(\sqrt{\frac{(p+f)\log M}{n}}\right) \tag{12}$$

as $n \to \infty$.

Proof. The proof uses results from large deviation theory. See [10]. □

Thm. 2 generalizes Thm. 1 to the case of sparse Kronecker structure. Comparison between the error expressions (4), (11) and (12) show that, by exploiting both Kronecker structure and sparsity, KGlasso can attain significantly lower estimation error than standard Glasso [3] and FF [1].

7. SIMULATION RESULTS

In this section, we compare the KGlasso algorithm with the flip-flop (FF) algorithm [1] that iteratively computes the ML solution. The Glasso algorithm implementation used was based on [8] with a stopping criterion determined by when the duality gap falls below a threshold of 10^{-5}.

To empirically evaluate performance, Monte Carlo simulations were used. The true matrices $\mathbf{X}_0 := \mathbf{A}_0^{-1}$ and $\mathbf{Y}_0 := \mathbf{B}_0^{-1}$ were unstructured randomly generated positive definite matrices based on an Erdös-Rényi graph model. Performance assessment was based on normalized Frobenius norm error in the covariance and precision matrix estimates. The normalized error was calculated using $\sqrt{\frac{1}{N_{MC}} \sum_{i=1}^{N_{MC}} \frac{\|\mathbf{\Sigma}_0 - \hat{\mathbf{\Sigma}}(i)\|_F^2}{\|\mathbf{\Sigma}_0\|_F^2}}$, where N_{MC} is the number of Monte Carlo runs and $\hat{\mathbf{\Sigma}}(i)$ is the covariance output from the ith trial run. [1]

[1] The same formula can be adapted to calculate the normalized error in the precision matrix $\hat{\mathbf{\Theta}}_0$.

In our implementation of KGlasso, the regularization parameters were chosen as follows. The initialization was $\mathbf{X}_{init} = \mathbf{I}_p$. The regularization parameters were selected as $\lambda_Y^{(1)} = c_y\sqrt{\frac{\log M}{np}}$, $\lambda_X^{(2)} = c_x\sqrt{\frac{\log M}{nf}} + \lambda_Y^{(1)}$, $\lambda_Y^{(2)} = \lambda_X^{(2)}$, $\lambda_X^{(3)} = \lambda_X^{(2)}$, etc. We set $c_x = c_y = 0.4$.

We consider the setting where $\mathbf{X}_0$ and $\mathbf{Y}_0$ are large sparse matrices of dimension $p = f = 60$ (see Fig. 1). The dimension of $\mathbf{\Theta}_0$ is $d = 3,600$, which is too large for standard Glasso to handle. Thus, it is not shown.

Figure 2 compares the root-mean squared error (RMSE) performance in precision and covariance matrices as a function of n. As expected, KGlasso outperforms FF [1] over the exhibited range of n for both the covariance and the inverse covariance estimation problem.

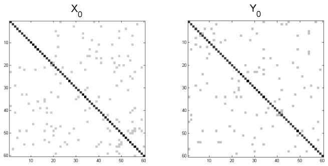

Fig. 1. Sparse Kronecker matrix representation. Left panel: left Kronecker factor. Right panel: right Kronecker factor.

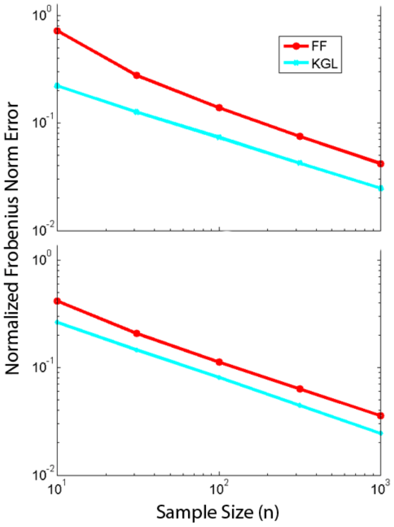

Fig. 2. Normalized RMSE performance for precision matrix (top) and covariance matrix (bottom) as a function of n. For $n = 10$, there is a 69% RMSE reduction for the precision matrix and 36% RMSE reduction for the covariance matrix when using KGlasso instead of FF.

8. CONCLUSION

We considered high-dimensional estimation of a Kronecker-decomposable covariance matrix given i.i.d. Gaussian samples. A ℓ_1-penalized likelihood approach was proposed for estimating the covariance matrix when the kronecker factors are sparse. This led to an iterative algorithm (KGlasso) that takes advantage of structure and sparsity. A tight MSE convergence rate was derived for KGlasso, showing significantly better MSE performance than standard Glasso and FF [1]. Simulations validated our theoretical predictions.

9. ACKNOWLEDGEMENT

The authors thank Prof. Mark Rudelson for very helpful discussions on large deviation theory.

10. REFERENCES

[1] K. Werner, M. Jansson, and P. Stoica, "On estimation of covariance matrices with Kronecker product structure," *IEEE Trans. Sig. Proc.*, vol. 56, no. 2, Feb. 2008.

[2] S. L. Lauritzen, *Graphical Models*, Oxford University Press, 1996.

[3] A. Rothman, P. Bickel, E. Levina, and J. Zhu, "Sparse permutation invariant covariance estimation," *Electronic Journal of Statistics*, vol. 2, pp. 494–515, 2008.

[4] S. Zhou, J. Lafferty, and L. Wasserman, "Time varying undirected graphs," *Machine Learning Journal*, vol. 80, pp. 295–319, 2010.

[5] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu, "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence," *Electronic Journal of Statistics*, vol. 5, pp. 935–980, 2011.

[6] N. Lu and D. Zimmerman, "On likelihood-based inference for a separable covariance matrix," Tech. Rep., Statistics and Actuarial Sc. Dept., Univ. Iowa, IA, 2004.

[7] 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*, March 2008.

[8] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[9] G. I. Allen and R. Tibshirani, "Transposable regularized covariance models with an application to missing data imputation," *The Annals of Applied Statistics*, vol. 4, no. 2, pp. 764–790, 2010.

[10] T. Tsiligkaridis, A. O. Hero III, and S. Zhou, "Convergence properties of kronecker graphical lasso algorithms," *arXiv:1204.0585v1*, 2012.