On Convergence of Kronecker Graphical Lasso Algorithms

Theodoros Tsiligkaridis *, *Student Member, IEEE*, Alfred O. Hero III, *Fellow, IEEE*, Shuheng Zhou, *Member, IEEE*

***Abstract*—This paper presents a thorough convergence analysis of Kronecker graphical lasso (KGLasso) algorithms for estimating the covariance of an i.i.d. Gaussian random sample under a sparse Kronecker-product covariance model. The KGlasso model, originally called the transposable regularized covariance model by Allen *et al* [1], implements a pair of ℓ_1 penalties on each Kronecker factor to enforce sparsity in the covariance estimator. The KGlasso algorithm generalizes Glasso, introduced by Yuan and Lin [2] and Banerjee *et al* [3], to estimate covariances having Kronecker product form. It also generalizes the unpenalized ML flip-flop (FF) algorithm of Werner *et al* [4] to estimation of sparse Kronecker factors. We establish high dimensional rates of convergence to the true covariance as both the number of samples and the number of variables go to infinity. Our results establish that KGlasso has significantly faster asymptotic convergence than Glasso and FF. Simulations are presented that validate the results of our analysis. For example, for a sparse $10,000 \times 10,000$ covariance matrix equal to the Kronecker product of two 100×100 matrices, the root mean squared error of the inverse covariance estimate using FF is 3.5 times larger than that obtainable using KGlasso.**

***Index Terms*—Sparsity, structured covariance estimation, penalized maximum likelihood, graphical lasso, direct product representation.**

I. Introduction

Covariance estimation is a problem of great interest in many different disciplines, including machine learning, signal processing, economics and bioinformatics. In many applications the number of variables is very large, e.g., in the tens or hundreds of thousands, leading to a number of covariance parameters that greatly exceeds the number of observations. To address this problem constraints are frequently imposed on the covariance to reduce the number of parameters in the model. For example, the Glasso model of Yuan and Lin [2] and Banerjee *et al* [3] imposes sparsity constraints on the covariance. The Kronecker product model of Werner *et al* [4] assumes that the covariance can be represented as the Kronecker product of two lower dimensional covariance matrices. The transposable regularized covariance model of Allen *et al* [1] imposes a combination of sparsity and Kronecker product form on the covariance. When there is no missing data, an extension of the alternating optimization algorithm of [4], called the flip flop (FF) algorithm, can be applied to estimate the parameters of this combined sparse and Kronecker product model. In this paper we call this algorithm the Kronecker Glasso (KGlasso) and we thoroughly analyze convergence of the algorithm in the high dimensional setting.

As in [4] we assume that there are pf variables whose covariance $\mathbf{\Sigma}_0$ has the separable positive definite Kronecker product representation:

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

where $\mathbf{A}_0$ is a $p \times p$ positive definite matrix and $\mathbf{B}_0$ is an $f \times f$ positive definite matrix. This model (1) is relevant to channel modeling for MIMO wireless communications, where $\mathbf{A}_0$ is a transmit covariance matrix and $\mathbf{B}_0$ is a receive covariance matrix [5]. The model is also relevant to other transposable models arising in recommendation systems like NetFlix and in gene expression analysis [1]. The Kronecker factorization (1) can easily be generalized to the k-fold case, where $\mathbf{\Sigma}_0 = \mathbf{A}_1 \otimes \mathbf{A}_2 \otimes \cdots \otimes \mathbf{A}_k$.

Under the assumption that the measurements are multivariate Gaussian with covariance having the Kronecker product form (1), the maximum likelihood (ML) estimator can be formulated [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 [6], [4]. As compared to the standard saturated (unstructured) covariance model, the number of unknown parameters in (1) is reduced from order $\Theta(p^2 f^2)$ to order $\Theta(p^2) + \Theta(f^2)$. This results in a significant reduction in the mean squared error (MSE) and the computational complexity of the maximum likelihood (ML) covariance estimator. This paper establishes that further reductions MSE are achievable when the Kronecker matrix factors are known to have sparse inverses, i.e., the measurements obey a sparse Kronecker structured Gaussian graphical model.

The graphical lasso (Glasso) estimator was originally proposed in [2], [3] for estimating a sparse inverse covariance, also called the precision matrix, under an i.i.d. Gaussian observation model. An algorithm for efficiently solving the nonsmooth optimization problem that arises in the Glasso estimator, based on ideas from [3], was proposed in [7]. Glasso has been applied to the time-varying coefficients setting in Zhou *et al* [8] using the kernel estimator for covariances at a target time. Rothman *et al* [9] derived high dimensional convergence rates for a slight variant of Glasso, i.e., only the off-diagonal entries of the estimated precision matrix were penalized using an ℓ_1-penalty. The high dimensional convergence rate of Glasso was established by Ravikumar *et al* [10]. This paper extends their analysis to the case that the covariance has Kronecker structure (1), showing that significantly higher rates of convergence are achievable.

The main contribution is the derivation of the high-dimensional MSE convergence rates for KGlasso as n, p and f go to infinity. When both Kronecker factors are sparse, it is shown that KGlasso *strictly* outperforms

FF and Glasso in terms of MSE convergence rate. More specifically, we show KGlasso achieves a convergence rate of $O_P\left(\frac{(p+f)\log\max(p,f,n)}{n}\right)$ and FF achieves a rate of $O_P\left(\frac{(p^2+f^2)\log\max(p,f,n)}{n}\right)$ as $n\to\infty$, while it is known [9], [8] that Glasso achieves a rate of $O_P\left(\frac{(pf+s)\log\max(p,f,n)}{n}\right)$, where s denotes the number of off-diagonal nonzero elements in the true precision matrix $\mathbf{\Theta}_0$. Simulations show that the performance improvements predicted by the high-dimensional analysis continue to hold for small sample size and moderate matrix dimension. For the example studied in Sec. VIII the empirical MSE of KGlasso is significantly lower than that of Glasso and FF for $p=f=100$ over the range of n from 10 to 100.

The starting point for the MSE convergence analysis is the large-sample analysis of the FF algorithm (Thm. 1 in [4]). The KGlasso convergence proof uses a large deviation inequality that shows that the dimension of one estimated Kronecker factor, say $\mathbf{A}$, acts as a multiplier on the number of independent samples when performing inference on the other factor $\mathbf{B}$. This result is then used to obtain optimal MSE rates in terms of Frobenius norm error between the KGlasso estimated matrix and the ground truth. The asymptotic MSE convergence analysis is useful since it can be used to guide the selection of sparsity regularization parameters and to determine minimum sample size requirements.

The outline of the paper is as follows. Section II introduces the notation that will be used throughout the paper. In Section III, the graphical lasso framework is introduced. Section IV uses this framework to introduce the KGlasso algorithm. Section V shows convergence of KGlasso and characterizes its limit points. The high dimensional MSE convergence rate derivation for the FF algorithm is included in Section VI. Section VII presents a high-dimensional MSE rate result that is used to establish the superiority of KGlasso as compared to FF and standard Glasso, under the sparse Kronecker product representation (1). Section VIII presents simulations that empirically validate the theoretical convergence rates obtained in Section VII.

II. 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 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 element 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 $\lambda_{min}(\mathbf{M}) = \min\lambda_i(\mathbf{M})$. Define the sparsity parameter associated with $\mathbf{M}$ as $s_M = \text{card}(\{(i_1,i_2) : [\mathbf{M}]_{i_1,i_2}\neq 0, i_1\neq i_2\})$. Let $\kappa(\mathbf{M}) := \frac{\lambda_{max}(\mathbf{M})}{\lambda_{min}(\mathbf{M})}$ denote the condition number of a symmetric matrix $\mathbf{M}$.

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 $S^p = \{\mathbf{A}\in\mathbb{R}^{p\times p} : \mathbf{A}=\mathbf{A}^T\}$, the set of symmetric positive semidefinite (psd) matrices S_+^p, and the set of symmetric positive definite (pd) matrices S_{++}^p. $\mathbf{I}_d$ is a $d\times d$ identity matrix. It can be shown that S_{++}^p is a convex set, but is not closed [11]. Note that S_{++}^p is simply the interior of the closed convex cone S_+^p.

Statistical convergence rates will be denoted by the $O_P(\cdot)$ notation, which is defined as follows. Consider a sequence of real random variables $\{X_n\}_{n\in\mathbb{N}}$ defined on a probability space $(\Omega,\mathcal{F},P)$ and a deterministic (positive) sequence of reals $\{b_n\}_{n\in\mathbb{N}}$. By $X_n = O_P(1)$ is meant: $\sup_{n\in\mathbb{N}}\Pr(|X_n|>K)\to 0$ as $K\to\infty$, where $\mathbf{X}_n$ is a sequence indexed by n, for fixed p,f. The notation $X_n = O_P(b_n)$ is equivalent to $\frac{X_n}{b_n} = O_P(1)$. By $X_n = o_p(1)$ is meant $\Pr(|X_n|>\epsilon)\to 0$ as $n\to\infty$ for any $\epsilon>0$. By $\lambda_n\asymp b_n$ is meant $c_1\leq\frac{\lambda_n}{b_n}\leq c_2$ for all n, where $c_1,c_2>0$ are absolute constants.

III. Graphical Lasso Framework

For simplicity, we assume the number of Kronecker components is $k=2$. Available are n i.i.d. multivariate Gaussian observations $\{\mathbf{z}_t\}_{t=1}^n$, where $\mathbf{z}_t\in\mathbb{R}^{pf}$, having zero-mean and covariance equal to $\mathbf{\Sigma}=\mathbf{A}_0\otimes\mathbf{B}_0$. Then, ignoring irrelevant constants, the log-likelihood $l(\mathbf{\Sigma})$ is:

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

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 [3], [7], [10] 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 obtained as 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 [7] 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 [9], [8]:

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

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 that achieved in the case of the standard sample covariance estimator ($\lambda = 0$):

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

IV. 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 an initial guess of $\mathbf{A}_0 = \mathbf{X}_0^{-1}$.

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

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

Although the objective function (6) is not jointly convex in $(\mathbf{X}, \mathbf{Y})$, it is biconvex. This motivates the flip-flop algorithm [4]. Adapting the notation from [4], 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 $\overline{\hat{\mathbf{S}}}_n = \mathbf{K}_{p,f}^T \hat{\mathbf{S}}_n \mathbf{K}_{p,f}$ (see Sec. II for definition of $K_{p,f}$). 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 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. This 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 minimize

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

where $\bar{\lambda}_X = \lambda_X/f$ and $\bar{\lambda}_Y = \lambda_Y/p$. 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 is the same objective function that was proposed in [1] when specialized to the case that there is no missing data.

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

$$\mathbf{G}(\mathbf{T}, \lambda) = \arg\min_{\mathbf{\Theta}\in S_{++}^d} \Big\{\text{tr}(\mathbf{\Theta}\mathbf{T}) - \log\det(\mathbf{\Theta}) + \lambda|\mathbf{\Theta}|_1\Big\}. \tag{10}$$

As compared to the $\mathcal{O}(p^3 f^3)$ computational complexity of Glasso, KGlasso has a computational complexity of only $\mathcal{O}(p^3 + f^3)$.

V. Convergence of KGlasso Iterations

In this section, we provide an alternative characterization of the KGlasso algorithm (Algorithm 1) and show the iterations converge pointwise to a critical point of the objective function. Under a mild condition on the starting point they will converge to a local minimum.

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
$\hat{\mathbf{\Theta}}_{KGlasso} \leftarrow \check{\mathbf{X}} \otimes \check{\mathbf{Y}}$

A. Block-Coordinate Reformulation of KGlasso

The following lemma shows that exploiting the property that the KGlasso algorithm is a block-coordinate optimization of the penalized objective function (9), each subproblem takes the form of standard Glasso applied on a compressed version of the SCM that is relevant for inference in each step.

Lemma 1. *The KGlasso objective function (9) has the following properties:*

1) *Assume $\bar{\lambda}_X, \bar{\lambda}_Y \geq 0$ and $\mathbf{X} \in S_{++}^p, \mathbf{Y} \in S_{++}^f$. When one argument of $J_\lambda(\mathbf{X}, \mathbf{Y})$ is fixed, the objective function (9) is convex in the other argument.*
2) *Assume $\hat{\mathbf{S}}_n$ is positive definite. Consider $J_\lambda(\mathbf{X}, \mathbf{Y})$ in (9) with matrix $\mathbf{X} \in S_{++}^p$ fixed. Then, the dual subproblem for minimizing $J_\lambda(\mathbf{X}, \mathbf{Y})$ over $\mathbf{Y}$ is:*

$$\max_{|\mathbf{W} - \frac{1}{p}\sum_{i,j=1}^{p} \mathbf{X}_{i,j}\hat{\mathbf{S}}_n(j,i)|_\infty \leq \lambda_Y} \log\det(\mathbf{W}) \tag{11}$$

where $\lambda_Y := \bar{\lambda}_Y/p$.
On the other hand, consider (9) with matrix $\mathbf{Y} \in S_{++}^f$ fixed. Then, the dual problem for minimizing $J_\lambda(\mathbf{X}, \mathbf{Y})$ over $\mathbf{X}$ is:

$$\max_{|\mathbf{Z} - \frac{1}{f}\sum_{k,l=1}^{f} \mathbf{Y}_{k,l}\overline{\hat{\mathbf{S}}_n}(l,k)|_\infty \leq \lambda_X} \log\det(\mathbf{Z}) \tag{12}$$

where $\overline{\hat{\mathbf{S}}_n} := \mathbf{K}_{p,f}^T \hat{\mathbf{S}}_n \mathbf{K}_{p,f}$ and $\lambda_X := \bar{\lambda}_X/f$.
3) *Strong duality holds for (11) and (12).*
4) *The solutions to (11) and (12) are positive definite.*

Proof: See Appendix. ∎

Since the dual subproblems (11) and (12) are maximizations of a strictly concave function over a closed convex set they have unique solution attaining the maximum. Lemma 1 is similar to the result obtained in [3], but with $(\frac{1}{p}\sum_{i,j=1}^{p} \mathbf{X}_{i,j}\hat{\mathbf{S}}_n(j,i), \lambda_Y)$ playing the role of $(\hat{\mathbf{S}}_n, \lambda)$, for the fixed $\mathbf{X}$ subproblem.

B. Limit Point Characterization of KGlasso

The following theorem establishes that KGlasso converges to a local minimum of the penalized likelihood function (2).

Theorem 1. *Assume $n > pf$. Then the KGlasso iterations converge to a critical point of the negative penalized likelihood function (9).*

Assuming $(\mathbf{X}^{(0)}, \mathbf{Y}^{(0)})$ is not a local maximum, the KGlasso iterations converge to a local minimum of the negative penalized likelihood function (9).

Proof: See [12]. Also see Prop. 2 in [1] for an alternative proof of the first part. ∎

The proof of Thm. 1 is built on several lemmas (see [12]). The main line of argument is as follows. For $n > pf$, the SCM is positive definite a.s., which implies that the objective function is bounded below. This can be used to show that the iterates generated by Algorithm 1 converge to a fixed point. Combining this result with the KKT optimality conditions and the strict descent property of the algorithm, we arrive to the result in Thm. 1.

VI. 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 $O_P\left(\sqrt{\frac{(p^2+f^2)\log M}{n}}\right)$. This result (see Thm. 2) will be compared to the statistical convergence rate of KGlasso (see Thm. 3) to establish that KGlasso has lower asymptotic MSE than FF. We make the following boundedness assumptions 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 $\mathbf{\Sigma}_{FF}(3) := \hat{\mathbf{A}}(\hat{\mathbf{B}}(\mathbf{A}_{init})) \otimes \hat{\mathbf{B}}(\hat{\mathbf{A}}(\hat{\mathbf{B}}(\mathbf{A}_{init})))$ denote the 3-step (noniterative) version of the flip-flop algorithm [4].

Theorem 2. *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,*

$$\|\mathbf{\Theta}_{FF}(3) - \mathbf{\Theta}_0\|_F = O_P\left(\sqrt{\frac{(p^2+f^2)\log M}{n}}\right) \qquad (13)$$

as $n \to \infty$.

Proof: See Appendix. ∎

Remark 1. *The sufficient conditions are symmetric with respect to p and f-i.e. for $f \geq p$, the corresponding conditions would become $f \log M \leq C'' n$ for some constant $C'' > 0$, and $n \geq \frac{f}{p} + 1$.*

For the special case of $p = f$, the sufficient conditions of Thm. 2 become $p \log M = O(n)$. The relation (13) indicates that the error is asymptotically bounded as long as n is of order $\Omega((p^2+f^2)\log M)$. The relation (13) specifies the rate of reduction of the estimation error for the three step FF algorithm ($k = 3$) [4]. This relation will also hold for the multi-step FF as long as the number of steps are finite. Note that (13) specifies a faster rate than that of the ordinary ML sample covariance matrix estimator (5).

VII. High Dimensional Consistency of KGlasso

Here a relation like (13) is established for KGlasso. Recall that a $p \times p$ matrix is called sparse if its number of nonzero elements is of order p. Recall $\lambda_X = \bar{\lambda}_X f$ and $\lambda_Y = \bar{\lambda}_Y p$, as in (9).

Theorem 3. *Assume $\mathbf{X}_0$ and $\mathbf{Y}_0$ are sparse. Let $\mathbf{A}_0, \mathbf{B}_0, \mathbf{A}_{init}$ satisfy Assumptions 1. Let $M = \max(p, f, n)$. Let $\lambda_Y^{(1)} \asymp \sqrt{\frac{\log M}{np}}$, and $\lambda_X^{(2)}, \lambda_Y^{(3)} \asymp \left(\frac{1}{\sqrt{p}} + \frac{1}{\sqrt{f}}\right)\sqrt{\frac{\log M}{n}}$. Then, if $\max\left(\frac{p}{f}, \frac{f}{p}\right)\log M = o(n)$,*

$$\|\mathbf{\Theta}_{KGlasso}(3) - \mathbf{\Theta}_0\|_F = O_P\left(\sqrt{\frac{(p+f)\log M}{n}}\right) \qquad (14)$$

as $n \to \infty$.

Proof: See Appendix. ∎

Theorem 3 offers a strict improvement over standard Glasso [9], [3] and generalizes Thm. 1 in [9] to the case of sparse Kronecker product structure. Thm. 3 generalizes Thm. 2 to the case of sparse Kronecker structure. Comparison between the error expressions (4), (13) and (14) show that, by exploiting both Kronecker structure and sparsity, KGlasso can attain significantly lower estimation error than standard Glasso [9] and FF [4]. To achieve accurate covariance estimation for the sparse Kronecker product model, the minimal sample size needed is $n = \Omega((p+f)\log M)$.

The minimal sample size required to achieve accurate covariance estimation is graphically depicted in Fig. 1 for the special case $p = f$. The regions below the lines are the MSE convergence regions-i.e., the MSE convergence rate goes to zero as p, n grow together to infinity at a certain growth rate controlled by these regions. It is shown that KGlasso allows the dimension p to grow almost linearly in n and still achieve accurate covariance estimation (see (14)) and thus, uniformly outperforms FF, Glasso and the naive SCM estimators in the case both Kronecker factors are sparse.

Although Thm. 3 shows a rate on the inverse covariance matrix, this asymptotic rate can be shown to hold for the covariance matrix as well (see proof of Thm. 3 in Appendix).

A. Discussion

Theorem 3 is established using the large deviation bound in Lemma 2. We provide some intuition on this bound below. Assume that $\mathbf{X}_{init} = \mathbf{X}_0$, or $\mathbf{A}_{init} = \mathbf{X}_{init}^{-1} = \mathbf{A}_0$. Define $\mathbf{W} = \mathbf{X}_0^{1/2} \otimes \mathbf{I}_p$ and $\tilde{\mathbf{z}}_t = \mathbf{W}\mathbf{z}_t$, with i.i.d. $\mathbf{z}_t \sim N(\mathbf{0}, \mathbf{A}_0 \otimes \mathbf{B}_0)$, $t = 1, \ldots, n$. Then, $\tilde{\mathbf{z}}_t$ has block-diagonal covariance

$$\mathrm{Cov}(\tilde{\mathbf{z}}_t) = \mathbf{I}_p \otimes \mathbf{B}_0.$$

When $\mathbf{W}$ is applied to the transformed $pf \times pf$ sample covariance matrix, $\hat{\mathbf{S}}_n^W := \mathbf{W}\hat{\mathbf{S}}_n\mathbf{W}^T$, the first step of KGlasso produces an iterate $\hat{\mathbf{Y}}_n^{(1)} = \mathbf{G}(\hat{\mathbf{B}}, \lambda_Y)$ with $\hat{\mathbf{B}} = \frac{1}{p}\sum_{i=1}^p \hat{\mathbf{S}}_n^W(i,i)$ (recall (8)). For suitable $\lambda_Y = \lambda_Y^{(1)}$, $\hat{\mathbf{Y}}_n^{(1)}$ converges to

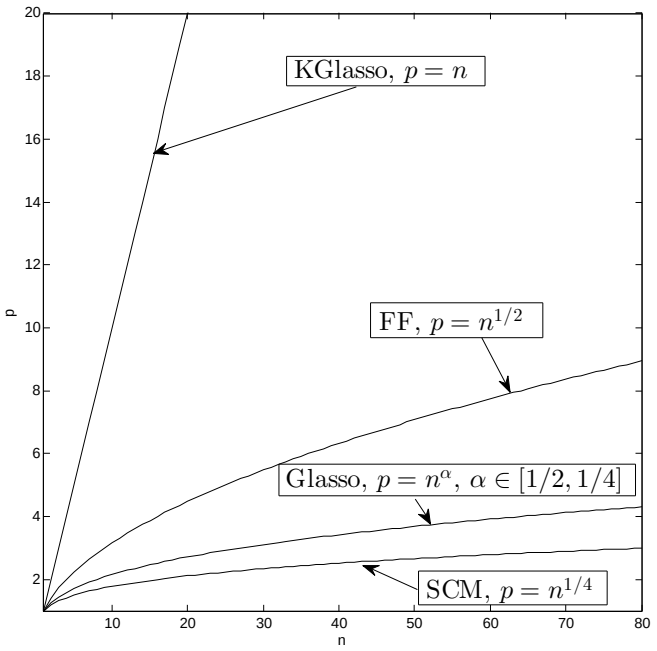

Fig. 1. Regions of convergence for KGlasso (below upper curve), FF (below second highest curve), Glasso (below third highest curve), and standard sample covariance matrix estimator (SCM) (bottom curve). These regions are obtained from the analytical expressions in equations (14), (13), (4) and (5), respectively. The simulation shown in Fig. 6 establishes that the FF algorithm indeed diverges when the parameters p and n fall inbetween the KGlasso and FF curves in the above figure.

$\mathbf{Y}_0$ with respect to maximal elementwise norm at a rate $O_P\left(\sqrt{\frac{\log M}{np}}\right)$. The convergence of $\hat{\mathbf{Y}}_n^{(1)}$ is easily established by applying the Chernoff bound and invoking the jointly Gaussian property of the measurements and the block diagonal structure of $\mathrm{Cov}(\tilde{\mathbf{z}}_t)$. Lemma 2 in the Appendix establishes that this rate holds even if $\mathbf{X}_{init} \neq \mathbf{X}_0$ in Assumption 1. In view of the rate of convergence of $\hat{\mathbf{Y}}^{(1)}$, to achieve a reduction in the MSE of $\mathbf{Y}$, either the sample size n or the dimension p must increase. Lemma 2 provides a tight bound that makes the dependence of the convergence rate explicit in p, f and n. Theorem 3 uses Lemma 2 to show that KGlasso converges to $\mathbf{X}_0 \otimes \mathbf{Y}_0$ with rate $O_P\left(\sqrt{\frac{(p+f)\log M}{n}}\right)$ with respect to Frobenius norm.

VIII. Simulation Results

In this section, we empirically validate the convergence rates established in previous sections using Monte Carlo simulation.

Each iteration of the KGlasso involves solving an ℓ_1 penalized covariance estimation problem of dimension 100×100 (Step 6 and Step 8 of KGlasso specified by Algorithm 1). To solve these small sparse covariance estimation problems we used the Glasso algorithm of Hsieh *et al* [13] where the Glasso stopping criterion was determined by monitoring when the duality gap falls below a threshold of 10^{-3}.

In each of the simulations the true covariance matrix factors $\mathbf{X}_0 = \mathbf{A}_0^{-1}$ and $\mathbf{Y}_0 = \mathbf{B}_0^{-1}$ were unstructured randomly generated positive definite matrices. First, p random nonzero elements were placed on the diagonal of a square $p \times p$ matrix C. Then, on average p nonzero elements were placed on the off-diagonal and symmetry was imposed. On average, a total of $3p$ elements were nonzero. The resulting matrix $\tilde{\mathbf{C}}$ was regularized to produce the sparse positive definite inverse covariance $\mathbf{Y}_0 = \tilde{\mathbf{C}} + \rho \mathbf{I}_f$, where $\rho = 0.5 - \lambda_{min}(\tilde{\mathbf{C}})$. A total of $N_{MC} = 50$ simulation runs were performed for each sample size n, where n ranged from 10 to 100. 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 $\hat{\mathbf{\Sigma}}(i)$ is the covariance estimate for the i-th simulation. The same formula was used to calculate the normalized error in the precision matrix $\hat{\mathbf{\Theta}}_0$. In the 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 $\lambda_Y^{(1)} = c_y \sqrt{\frac{\log M}{np}}$, $\lambda_X^{(2)} = c_x \sqrt{\frac{\log M}{nf}} + \lambda_Y^{(1)}$, $\lambda_Y^{(2)} = \lambda_X^{(2)}$, $\lambda_X^{(3)} = \lambda_X^{(2)}$ and so on. We set $c_x = c_y = 0.4$.

We considered the setting where $\mathbf{X}_0$ and $\mathbf{Y}_0$ are large sparse matrices of dimension $p = f = 100$ (see Fig. 2) yeilding a covariance matrix $\mathbf{\Theta}_0$ of dimension $10,000 \times 10,000$. This dimension was too large for implementation of Glasso even when implemented using the state-of-the-art algorithm by Hsieh *et al* [13]. Figures 3 and 4 compare the root-mean squared error (RMSE) performance in precision and covariance matrices as a function of n. As expected, KGlasso outperforms FF over the range of n for both covariance and inverse covariance estimation problems. KGlasso outperforms FF in the small-sample regime since it exploits sparsity in addition to Kronecker structure.

For $n = 10$, there is a 72% (≈ 5.53 dB) RMSE reduction for the precision matrix and 49% RMSE reduction for the covariance matrix when using KGlasso instead of FF. For $n = 100$, there is a 51% (≈ 3.10 dB) RMSE reduction for the precision matrix and 41% RMSE reduction for the covariance matrix. For the small sample regime, there is approximately a 5.53 dB reduction for the precision matrix, which is a significant performance gain.

A. Empirical Rate Comparison

Next, we illustrate the rates obtained in for the dimension setting $p(n) = f(n) = \lceil 8n^{\alpha} \rceil$, where $\alpha \in \{0.1, 0.2, 0.3\}$. According to the theory developed, for large n, the MSE converges to zero at a certain convergence rate. The predicted rates of FF and KGlasso are fitted on top of the empirical MSE curves by ensuring intersection at $n = 1000$. Fig. 5 shows that the empirical rates match the predicted rates well.

We also show a borderline case $p = f = \lceil n^{0.6} \rceil$. In this case, according to Thm. 2 and Thm. 3, the FF diverges (MSE increases in n), while the KGlasso converges (MSE decreases in n). This is illustrated in Fig. 6. Our predicted rates are plotted on top of the empirical curves.

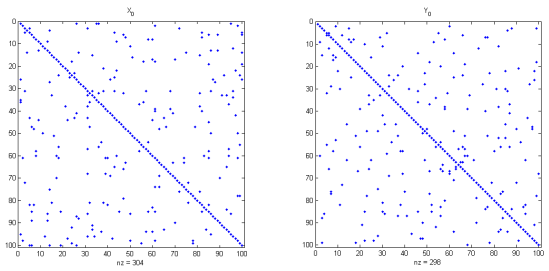

Fig. 2. Sparse Kronecker matrix representation. Left panel: left Kronecker factor. Right panel: right Kronecker factor. As the Kronecker-product covariance matrix is of dimension $10,000 \times 10,000$ standard Glasso is not practically implementable for this example.

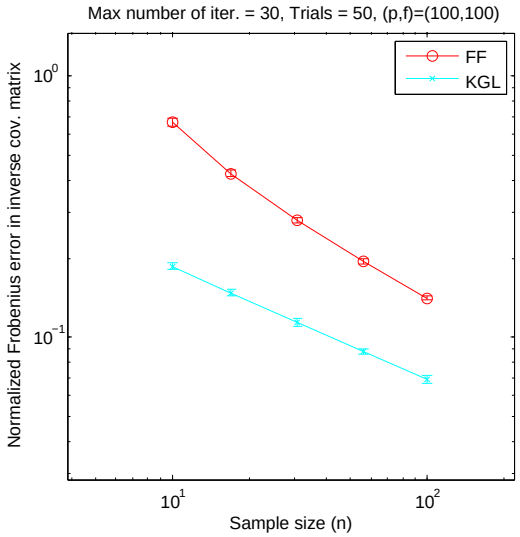

Fig. 3. Normalized RMSE performance for precision matrix as a function of sample size n. KGlasso (Kronecker graphical lasso) uniformly outperforms FF (flip-flop) algorithm for all n. Here, $p = f = 100$ and $N_{MC} = 50$. The error bars are centered around the mean with $\pm$ one standard deviation. For $n = 10$, there is a 72% RMSE reduction.

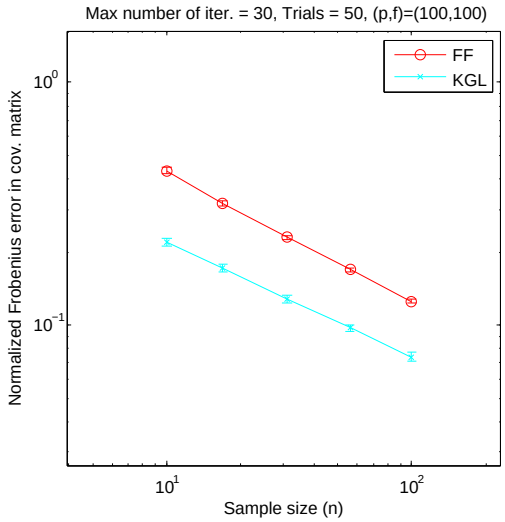

Fig. 4. Normalized RMSE performance for covariance matrix as a function of sample size n. KGlasso (Kronecker graphical lasso) uniformly outperforms FF (flip-flop) algorithm for all n. Here, $p = f = 100$ and $N_{MC} = 50$. The error bars are centered around the mean with $\pm$ one standard deviation. For $n = 10$, there is a 49% RMSE reduction.

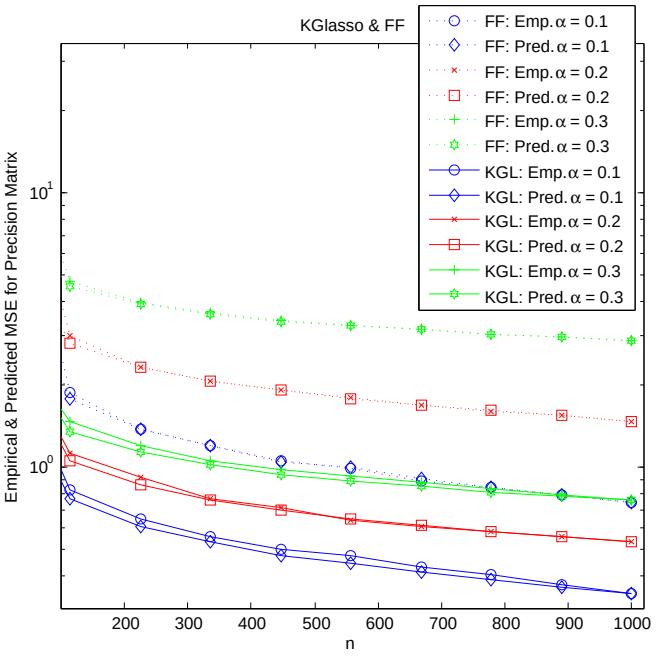

Fig. 5. Precision Matrix MSE convergence as a function of sample size n for FF and KGlasso. The dimensions of the Kronecker factor matrices grow as a function of n as: $p(n) = f(n) = \lceil 8 \cdot n^{\alpha} \rceil$. The true Kronecker factors were set to identity (so their inverses are fully sparse). The predicted MSE curves according to Thm. 2 and Thm. 3 are also shown. For both KGlasso and FF, the predicted MSE matches the empirical MSE well, thus verifying the rate expressions (13) and (14).

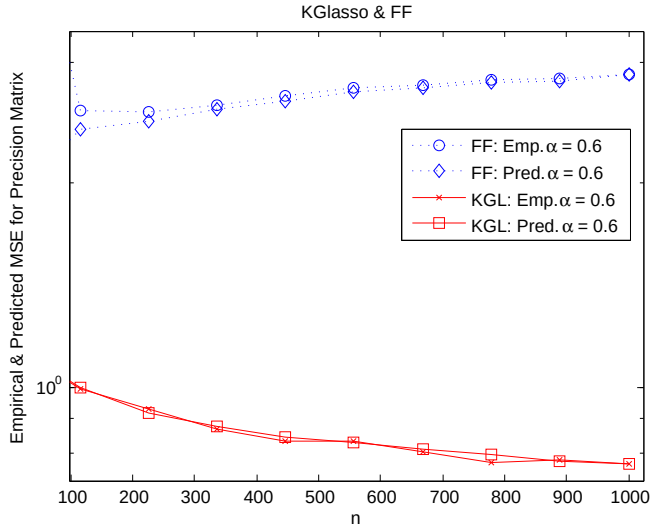

Fig. 6. Precision Matrix MSE as a function of sample size n for FF and KGlasso. The dimensions of the Kronecker factor matrices grow as a function of n as: $p(n) = f(n) = \lceil n^{0.6} \rceil$. The true Kronecker factors were set to identity (so their inverses are fully sparse). The predicted MSE curves according to Thm. 2 and Thm. 3 are also shown. As predicted by our theory, and by the predicted convergent regions of (n, p) for FF and KGlasso in Fig. 1, the MSE of the FF diverges while the MSE of the KGlasso converges as n increases.

IX. Conclusion

We established high dimensional consistency for Kronecker Glasso algorithms that use iterative ℓ_1-penalized likelihood optimization that exploit both Kronecker structure and sparsity of the covariance. A tight MSE convergence rate was derived for KGlasso, showing significantly better MSE performance than standard Glasso [9], [3] and FF [4]. Simulations validated our theoretical predictions.

Acknowledgement

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

Appendix A
Proof of Lemma 1

Proof:

1) Without loss of generality, fix $\mathbf{Y} \in S^f_{++}$. The function $\text{tr}((\mathbf{X} \otimes \mathbf{Y})\hat{\mathbf{S}}_n)$ is linear in $\mathbf{X}$. The function $g(\mathbf{X}_1) := -\log\det(\mathbf{X}_1)$ is a convex function in $\mathbf{X}_1$ over the set S^p_{++} [11]. The triangle inequality implies $|\cdot|_1$ is convex. Finally, the sum of convex functions is convex. The set S^p_{++} is a convex set for any $p \in \mathbb{N}$.
2) By symmetry we only need prove that (12) is the dual of $\min_{\mathbf{Y}\in S^f_{++}} J_\lambda(\mathbf{X}, \mathbf{Y})$. By standard duality relations between ℓ_1 and ℓ_∞ norms [11] and symmetry of $\mathbf{Y}$ [1]:

$$|\mathbf{Y}|_1 = \max_{\mathbf{U}\in S^f : |\mathbf{U}|_\infty \leq 1} \text{tr}(\mathbf{Y}\mathbf{U})$$

Using this in (9) and invoking the saddlepoint inequality:

$$\begin{aligned}&\min_{\mathbf{Y}\in S^f_{++}} \text{tr}((\mathbf{X}\otimes\mathbf{Y})\hat{\mathbf{S}}_n) - p\log\det(\mathbf{Y}) + p\lambda_Y|\mathbf{Y}|_1 \\ &\geq \max_{|\mathbf{U}|_\infty\leq\lambda_Y} \min_{\mathbf{Y}\in S^f_{++}} \Big\{\text{tr}((\mathbf{X}\otimes\mathbf{Y})\hat{\mathbf{S}}_n) - p\log\det(\mathbf{Y}) \\ &\qquad + p\text{tr}(\mathbf{Y}\mathbf{U})\Big\} \qquad (15)\end{aligned}$$

When the equality in (15) is achieved, $(\mathbf{U}, \mathbf{Y})$ is a saddlepoint and the duality gap is zero. Rewrite the objective function, denoted $\tilde{J}_\lambda(\cdot,\cdot)$, in the minimax operation (15):

$$\tilde{J}_\lambda(\mathbf{X}, \mathbf{Y}) := \text{tr}((\mathbf{X}\otimes\mathbf{Y})(\hat{\mathbf{S}}_n + \tilde{\mathbf{U}}(\mathbf{X}))) - p\log\det(\mathbf{Y})$$

where $\tilde{\mathbf{U}}(\mathbf{X}) = p\frac{\mathbf{I}_p\otimes\mathbf{U}}{\text{tr}(\mathbf{X})}$. Define $\mathbf{M} = \hat{\mathbf{S}}_n + \tilde{\mathbf{U}}(\mathbf{X})$. To evaluate $\min_{\mathbf{Y}\in S^f_{++}} \tilde{J}_\lambda(\mathbf{X}, \mathbf{Y})$ in (15), we invoke the KKT conditions to obtain the solution $\mathbf{Y} = \left(\frac{1}{p}\sum_{i,j=1}^p \mathbf{X}_{i,j}\mathbf{M}(j,i)\right)^{-1}$. Define $\mathbf{W} = \mathbf{Y}^{-1}$ as the dual space variable. Using this in (15):

$$\max_{|\mathbf{W} - \frac{1}{p}\sum_{i,j=1}^p \mathbf{X}_{i,j}\hat{\mathbf{S}}_n(j,i)|_\infty \leq \lambda_Y} p\log\det(\mathbf{W}) + pf \quad (16)$$

where the constraint set was obtained in terms of $\mathbf{W}$ by observing that $\tilde{\mathbf{U}}(\mathbf{X})(j,i) = \frac{p\mathbf{U}}{\text{tr}(\mathbf{X})} I(j=i)$, and $I(\cdot)$ is the indicator function. It is evident that (16) is equivalent to (11).

[1]The maximum is attained at $\mathbf{U}_{i,j} = \frac{\mathbf{Y}_{i,j}}{|\mathbf{Y}_{i,j}|}$ for $\mathbf{Y}_{i,j} \neq 0$ and at $\mathbf{U}_{i,j} = 0$ for $\mathbf{Y}_{i,j} = 0$.

3) It suffices to verify that the duality induced by the saddle point formulation is equivalent to Lagrangian duality (see Section 5.4 in [11]). Slater's constraint qualification (see Section 5.3.2 in [11]) trivially holds for the convex problem $\min_{\mathbf{Y}\in S^f_{++}} \tilde{J}_\lambda(\mathbf{X}, \mathbf{Y})$, and thus for the corresponding convex problem $\min_{\mathbf{Y}\in S^f_{++}} J_\lambda(\mathbf{X}, \mathbf{Y})$. Since the objective function of each dual problem has an optimal objective that is bounded below, Slater's constraint qualification also implies that the dual optimal solution is attained.
4) From [4], it follows that if $\hat{\mathbf{S}}_n$ is p.d., each "compression step" (see lines 6 and 8 in Algorithm 1) yields a p.d. matrix. Combining this with the positive definiteness of the Glasso estimator [3], we conclude that the first subiteration of KGlasso yields a p.d. matrix. A simple induction, combined with the fact that the Kronecker product of p.d. matrices is p.d., establishes that (11) and (12) are p.d.

■

Appendix B
Lemma 2

The following lemma will be used in the proof of Thm. 2 and Thm. 3.

Lemma 2. *Let $\mathbf{X}$ be a $p \times p$ data-independent matrix. Define the linear operator $\mathbf{T}$ as $\mathbf{T}(\mathbf{X}) = \hat{\mathbf{B}}(\mathbf{X}^{-1})$, where $\hat{\mathbf{B}}(\cdot)$ is defined in (8). Assume $\max_k[\mathbf{B}_0]_{k,k}, \|\mathbf{X}\|_2, \|\mathbf{A}_0\|_2$ are uniformly bounded constants as $p, f \to \infty$. Define $\mathbf{B}_* := \frac{tr(\mathbf{X}\mathbf{A}_0)}{p}\mathbf{B}_0$. Let $c, \tau > 0$. Define $\psi(u) = \sum_{m=0}^\infty \frac{(2m+2)!!}{m!} u^m$ [2]. Let $\bar{C} := \frac{4(2+\tau)^2\max(2,c)}{\psi(\frac{1}{2+\tau})} < \frac{np}{\log(\max(f,n))}$ [3]. Then, with probability $1 - \frac{2}{\max(f,n)^c}$,*

$$|\mathbf{T}(\mathbf{X}) - \mathbf{B}_*|_\infty \leq \overline{k}\cdot\sqrt{4\psi(\frac{1}{2+\tau})\max(2,c)}\sqrt{\frac{\log(\max(f,n))}{np}}$$

where $\overline{k} = \max_k[\mathbf{B}_0]_{k,k} \cdot \|\mathbf{X}\|_2\|\mathbf{A}_0\|_2$.

Remark 2. *Choosing $c \leq 2$ in Lemma 2, the best relative constant is obtained by taking τ to infinity, which yields $\sqrt{4\psi(\frac{1}{2+\tau})\max(2,c)} \to 4$.*

Remark 3. *For the case of symmetric matrices $\mathbf{X} \in S^p$, the constant $\overline{k}$ can be improved to $\max_k[\mathbf{B}_0]_{k,k} \cdot \|\mathbf{X}\mathbf{A}_0\|_2$.*

Proof: This proof is based on a large-deviation theory argument. Fix $(k,l) \in \{1,\ldots,f\}^2$. Note that $\text{E}[\mathbf{T}(\mathbf{X})] = \mathbf{B}_*$. First we bound the upper tail probability on the difference $\mathbf{T}(\mathbf{X}) - \mathbf{B}_*$ and then we turn to the lower tail probability.

[2]The double factorial notation is defined as

$$m!! = \begin{cases} m\cdot(m-2)\cdot\cdots\cdot 3\cdot 1 & \text{if } m>0 \text{ is odd} \\ m\cdot(m-2)\cdot\cdots\cdot 4\cdot 2 & \text{if } m>0 \text{ is even} \\ 1 & \text{if } m=-1 \text{ or } m=0 \end{cases}$$

.

[3]If $p = f = n^{c'}$ for some $c' > 0$, this condition will hold for n large enough.

Bounding the upper tail by using Markov's inequality, we have

$$\begin{aligned}
&\Pr\left([\mathbf{T}(\mathbf{X})]_{k,l} - [\mathbf{B}_*]_{k,l} > \epsilon\right) \\
&= \Pr\left(\frac{1}{p}\sum_{i,j=1}^{p} \mathbf{X}_{i,j}[\hat{\mathbf{S}}_n(j,i)]_{k,l} - \frac{\mathrm{tr}(\mathbf{X}\mathbf{A}_0)}{p}[\mathbf{B}_0]_{k,l} > \epsilon\right) \\
&= \Pr\Big(\exp\{t\sum_{m=1}^{n}\sum_{i,j=1}^{p}\mathbf{X}_{i,j}\Big([\mathbf{z}_m]_{(i-1)f+k}[\mathbf{z}_m]_{(j-1)f+l} \\
&\qquad - [\mathbf{A}_0]_{i,j}[\mathbf{B}_0]_{k,l}\Big)\} > \exp\{tnp\epsilon\}\Big) \\
&\leq e^{-tnp\epsilon}\Big(\mathrm{E}\Big[\exp\Big\{t\tilde{Y}^{(k,l)}\Big\}\Big]\Big)^n \qquad (17)
\end{aligned}$$

where we used the i.i.d. property of the data in (17) and $\tilde{Y}^{(k,l)} := \sum_{i,j=1}^{p}\mathbf{X}_{i,j}([\mathbf{z}]_{(i-1)f+k}[\mathbf{z}]_{(j-1)f+l} - [\mathbf{A}_0]_{i,j}[\mathbf{B}_0]_{k,l})$. Define $p^2 \times 1$ random vector $\mathbf{z}^{(k,l)}$ as $[\mathbf{z}^{(k,l)}]_{(i-1)p+j} := [\mathbf{z}]_{(i-1)f+k}[\mathbf{z}]_{(j-1)f+l} - [\mathbf{A}_0]_{i,j}[\mathbf{B}_0]_{k,l}$ for $1 \leq i,j \leq p$. Clearly, this random vector is zero mean. The expectation term inside the parentheses in (17) is the MGF of the random variable $\tilde{Y}^{(k,l)} = \mathrm{vec}(\mathbf{X})^T\mathbf{z}^{(k,l)}$. For notational simplicity, let $\tilde{\phi}_Y(t) = \mathrm{E}[e^{tY}]$ denote the MGF of a random vector Y.

Performing a second order Taylor expansion on $\tilde{\phi}_{\tilde{Y}^{(k,l)}}$ about the origin, we obtain:

$$\tilde{\phi}_{\tilde{Y}^{(k,l)}}(t) = \tilde{\phi}_{\tilde{Y}^{(k,l)}}(0) + \frac{d\tilde{\phi}_{\tilde{Y}^{(k,l)}}(0)}{dt}t + \frac{1}{2}\frac{d^2\tilde{\phi}_{\tilde{Y}^{(k,l)}}(\delta t)}{dt^2}t^2$$

for some $\delta \in [0,1]$. Trivially, $\tilde{\phi}_{\tilde{Y}^{(k,l)}}(0) = 1$ and $\frac{d\tilde{\phi}_{\tilde{Y}^{(k,l)}}(0)}{dt} = \mathrm{E}[\mathrm{vec}(\mathbf{X})^T\mathbf{z}^{(k,l)}] = 0$. Using the linearity of the expectation operator, we have:

$$\begin{aligned}
\frac{d^2\tilde{\phi}_{\tilde{Y}^{(k,l)}}(\delta t)}{dt^2} &= \mathrm{E}[(\tilde{Y}^{(k,l)})^2 e^{t\delta\tilde{Y}^{(k,l)}}] \\
&= \sum_{m=0}^{\infty}\frac{(\delta t)^m}{m!}\mathrm{E}[(\mathrm{vec}(\mathbf{X})^T\mathbf{z}^{(k,l)})^{m+2}]
\end{aligned}$$

Using the elementary inequality $1 + y \leq e^y$ for $y > -1$, and after some algebra, we have:

$$n\ln(\tilde{\phi}_{\tilde{Y}^{(k,l)}}(t)) \leq \frac{n}{2}t^2\sum_{m=0}^{\infty}T_m(t) \qquad (18)$$

where $T_m(t) := \frac{(t\delta)^m}{m!}\mathrm{E}[(\mathrm{vec}(\mathbf{X})^T\mathbf{z}^{(k,l)})^{m+2}]$. Note that

$$\begin{aligned}
t^2T_m(t) &\leq \frac{t^{m+2}}{m!}\mathrm{E}\Big[\Big(\sum_{i,j=1}^{p}\mathbf{X}_{i,j}([\mathbf{z}]_{(i-1)f+k}[\mathbf{z}]_{(j-1)f+l} \\
&\quad - [\mathbf{A}_0]_{i,j}[\mathbf{B}_0]_{k,l})\Big)^{m+2}\Big] \\
&= \frac{t^{m+2}}{m!}\sum_{i_1,j_1=1}^{p}\cdots\sum_{i_{m+2},j_{m+2}=1}^{p}\mathbf{X}_{i_1,j_1}\cdots\mathbf{X}_{i_{m+2},j_{m+2}} \\
&\quad \times \mathrm{E}\Big[\prod_{\alpha=1}^{m+2}\Big([\mathbf{z}]_{(i_\alpha-1)f+k}[\mathbf{z}]_{(j_\alpha-1)f+l} - [\mathbf{A}_0]_{i_\alpha,j_\alpha}[\mathbf{B}_0]_{k,l}\Big)\Big] \\
&\leq \frac{t^{m+2}}{m!}(2m+2)!!\cdot p\left(\max_k[\mathbf{B}_0]_{k,k}\|\mathbf{X}\|_2\|\mathbf{A}_0\|_2\right)^{m+2} \qquad (19) \\
&= \frac{(2m+2)!!}{m!}(t\bar{k})^{m+2}p
\end{aligned}$$

where (19) follows from Isserlis' formula [12]. Also, we defined $\bar{k} = \max_k[\mathbf{B}_0]_{k,k}\|\mathbf{X}\|_2\|\mathbf{A}_0\|_2$. Summing the result over m, and letting $u := t\bar{k} > 0$, $a_m(u) := \frac{(2m+2)!!}{m!}u^m$, $\psi(u) = \sum_{m=0}^{\infty}a_m(u)$, we obtain:

$$t^2\sum_{m=0}^{\infty}T_m(t) \leq pu^2\psi(u)\Big|_{u=t\bar{k}} \qquad (20)$$

By the ratio test [14], the infinite series $\sum_{m=0}^{\infty}a_m(u)$ converges if $u < 1/2$. Using (20) in (18), and the result in (17), we obtain the exponential bound:

$$\Pr([\mathbf{T}(\mathbf{X})]_{k,l} - [\mathbf{B}_*]_{k,l} > \epsilon) \leq \exp\left\{-tnp\epsilon + \frac{np(t\bar{k})^2}{2}\psi(t\bar{k})\right\}$$

Let $t < \frac{1}{(2+\tau)\bar{k}}$ and $\epsilon < \frac{1}{2+\tau}\psi(\frac{1}{2+\tau})\bar{k} < \infty$. By the monotonicity of $\psi(\cdot)$, we have:

$$\Pr([\mathbf{T}(\mathbf{X})]_{k,l} - [\mathbf{B}_*]_{k,l} > \epsilon) \leq \exp\left\{-tnp\epsilon + \frac{npt^2\bar{k}^2}{2}\psi(\frac{1}{2+\tau})\right\} \qquad (21)$$

Optimizing (21) over t, we obtain $t^* = \frac{\epsilon}{\bar{k}^2\psi(\frac{1}{2+\tau})}$. Clearly, $t^* < \frac{1}{(2+\tau)\bar{k}}$. Plugging this into (21) and letting $C := \frac{1}{2\bar{k}^2\psi(\frac{1}{2+\tau})}$ [4], we obtain for all $\epsilon < \frac{1}{2+\tau}\psi(\frac{1}{2+\tau})\bar{k}$:

$$\Pr([\mathbf{T}(\mathbf{X})]_{k,l} - [\mathbf{B}_*]_{k,l} > \epsilon) \leq e^{-np\epsilon^2C} \qquad (22)$$

From (22) and a similar lower tail bound, we conclude that for all $\epsilon < \frac{1}{2+\tau}\psi(\frac{1}{2+\tau})\bar{k}$:

$$\Pr(|[\mathbf{T}(\mathbf{X})]_{k,l} - \mathrm{E}[[\mathbf{T}(\mathbf{X})]_{k,l}]| > \epsilon) \leq 2e^{-np\epsilon^2C}$$

The union bound over $(k,l) \in \{1,\ldots,f\}^2$ completes the proof. This bound can be re-expressed as in the statement of Lemma 2 (see [12] for more details). ■

Appendix C
Proposition 1

Proposition 1. *Let $\mathbf{S}_{p,f,n}$ be a $d' \times d'$ (where $d' = p$ or $d' = f$) random matrix such that with probability $1 - \frac{2}{n^2}$, $|\mathbf{S}_{p,f,n} - \mathbf{\Sigma}_*|_\infty \leq r_{p,f,n}$. Assume $\mathbf{\Sigma}_* \in S_{++}^{d'}$ has uniformly bounded spectrum as $p, f \to \infty$ (analog to Assumption 1). Choose $\lambda_{p,f,n} = c \cdot r_{p,f,n}$ for some absolute constant $c > 0$. Consider the Glasso operator $\mathbf{G}(\cdot,\cdot)$ defined in (10). Let $s = s_{\mathbf{\Theta}_*}$ be the sparsity parameter associated with $\mathbf{\Theta}_* := \mathbf{\Sigma}_*^{-1}$. Assume $\sqrt{d'+s}\cdot r_{p,f,n} = o(1)$. Then, with probability $1 - \frac{2}{n^2}$,*

$$\|\mathbf{G}(\mathbf{S}_{p,f,n}, \lambda_{p,f,n}) - \mathbf{\Theta}_*\|_F \leq \frac{2\sqrt{2}(1+c)}{\lambda_{min}(\mathbf{\Sigma}_*)^2}\sqrt{d'+s}\cdot r_{p,f,n}$$

as $p, f, n \to \infty$.

Proof: The proof follows from a slight modification of Thm. 1 in [9], or Thm. 3 in [8]. This modification is due to the different $r_{p,f,n}$. ■

[4] Since $\psi(\frac{1}{2+\tau})$ is finite, $C > 0$ is finite.

APPENDIX D
PROOF OF THEOREM 2

Proof: As in the proof of Thm. 1 in [4], let $\mathbf{B}_* = \frac{\text{tr}(\mathbf{A}_0\mathbf{A}_{init}^{-1})}{p}\mathbf{B}_0$ and $\mathbf{A}_* = (\frac{\text{tr}(\mathbf{A}_0\mathbf{A}_{init}^{-1})}{p})^{-1}\mathbf{A}_0$. Note that Assumption 1 implies that $\|\mathbf{B}_*\|_2 = \Theta(1)$ and $\|\mathbf{A}_*\|_2 = \Theta(1)$ as $p, f \to \infty$. For conciseness, the statement "with probability $1 - \frac{2}{n^2}$" will be abbreviated as "w.h.p."-i.e., with high probability.

For concreteness, we first present the result for $k = 2$ iterations. Then, we generalize the analysis to all finite flip-flop iterations by induction.

The growth assumptions in the theorem imply

$$\max\left\{p, f, \frac{f^2}{p}, \left(\frac{\sqrt{pf} + f\sqrt{\frac{f}{p}} + p\sqrt{\frac{p}{f}}}{p+f}\right)^2\right\}\log M \leq C'n \tag{23}$$

for some constant $C' > 0$ large enough [5]. In fact, the growth assumption in the theorem statement can be relaxed to (23).

Define intermediate error matrices:

$$\tilde{\mathbf{B}}^0 = \hat{\mathbf{B}}(\mathbf{A}_{init}) - \mathbf{B}_*$$
$$\tilde{\mathbf{A}}^1 = \hat{\mathbf{A}}(\hat{\mathbf{B}}(\mathbf{A}_{init})) - \mathbf{A}_*$$

Define $\mathbf{Y}_* = \mathbf{B}_*^{-1}$ and $\mathbf{X}_* = \mathbf{A}_*^{-1}$. Also, define:

$$\mathbf{Y}_1 = \hat{\mathbf{B}}(\mathbf{A}_{init})^{-1}$$
$$\mathbf{X}_2 = \hat{\mathbf{A}}(\hat{\mathbf{B}}(\mathbf{A}_{init}))^{-1}$$

These inverses exist if $n \geq \max(\frac{p}{f}, \frac{f}{p}) + 1$ (see [15]). Define the error $\tilde{\mathbf{\Sigma}}_{FF}(k) = \mathbf{\Sigma}_{FF}(k) - \mathbf{\Sigma}_0$ for $k \geq 2$. For notational simplicity, let $\mathbf{B}_0^{max} := \max_k[\mathbf{B}_0]_{k,k}$ and $\mathbf{A}_0^{max} := \max_i[\mathbf{A}_0]_{i,i}$, $\psi_\tau := \psi(\frac{1}{2+\tau})$, where $\psi(\cdot)$ is defined in Lemma 2.

Lemma 2 implies that for

$$n > \frac{8(2+\tau)^2}{\psi_\tau}\log M \tag{24}$$

then with probability $1 - \frac{2}{n^2}$, we have:

$$\|\tilde{\mathbf{B}}^0\|_F \leq 2\sqrt{2\psi_\tau}\mathbf{B}_0^{max}\|\mathbf{A}_{init}^{-1}\mathbf{A}_0\|_2 f\sqrt{\frac{\log M}{np}} \tag{25}$$

As in the proof of Thm. 1 in [4], we vectorize the operations (7) and (8):

$$\text{vec}(\hat{\mathbf{A}}(\mathbf{B})) = \frac{1}{f}\hat{\mathbf{R}}_A\text{vec}(\mathbf{B}^{-1})$$
$$\text{vec}(\hat{\mathbf{B}}(\mathbf{A})) = \frac{1}{p}\hat{\mathbf{R}}_B\text{vec}(\mathbf{A}^{-1})$$

where $\hat{\mathbf{R}}_A$ and $\hat{\mathbf{R}}_B$ are permuted versions of the sample covariance matrix [4].

Let $\epsilon' > 1$. Note that from (25), for

$$n \geq (\epsilon' 2\sqrt{2\psi_\tau}\mathbf{B}_0^{max}\|\mathbf{A}_{init}^{-1}\mathbf{A}_0\|_2)^2 f^2 p^{-1}\log M \tag{26}$$

[5]This constant is independent of p, f, n, but may depend on the constants in Assumption 1.

with probability $1 - \frac{2}{n^2}$,

$$\lambda_{min}(\hat{\mathbf{B}}(\mathbf{A}_{init})) = \lambda_{min}(\tilde{\mathbf{B}}^0 + \mathbf{B}_*) \geq \lambda_{min}(\mathbf{B}_*) - \|\tilde{\mathbf{B}}^0\|_2$$
$$\geq \lambda_{min}(\mathbf{B}_*) - \|\tilde{\mathbf{B}}^0\|_F \geq \left(1 - \frac{1}{\epsilon'}\right)\lambda_{min}(\mathbf{B}_*)$$

Thus, w.h.p.,

$$\|\mathbf{Y}_1 - \mathbf{Y}_*\|_F = \|\mathbf{Y}_1(\hat{\mathbf{B}}(\mathbf{A}_{init}) - \mathbf{B}_*)\mathbf{Y}_*\|_F$$
$$\leq \|\mathbf{Y}_1\|_2\|\mathbf{Y}_*\|_2\|\tilde{\mathbf{B}}^0\|_F = \frac{\|\tilde{\mathbf{B}}^0\|_F}{\lambda_{min}(\mathbf{B}_*)\lambda_{min}(\hat{\mathbf{B}}(\mathbf{A}_{init}))}$$
$$\leq \left(1 - \frac{1}{\epsilon'}\right)^{-1}\|\mathbf{Y}_*\|_2^2 2\sqrt{2\psi_\tau}\mathbf{B}_0^{max}\|\mathbf{A}_{init}^{-1}\mathbf{A}_0\|_2$$
$$\times f p^{-1/2}\sqrt{\frac{\log M}{n}} \tag{27}$$

Using Lemma 2, for

$$n > \frac{8(2+\tau)^2}{\psi_\tau}\log M \tag{28}$$

then, w.h.p.,

$$\|\tilde{\mathbf{R}}_A\|_2 = \sup_{\|\mathbf{v}\|_2=1}\|\tilde{\mathbf{R}}_A\mathbf{v}\|_2 \leq p\sup_{\|\mathbf{v}\|_2=1}\|\hat{\mathbf{R}}_A\mathbf{v} - \mathbf{R}_A\mathbf{v}\|_\infty$$
$$= pf\sup_{\|\mathbf{v}\|_2=1}\|\frac{1}{f}\hat{\mathbf{R}}_A\mathbf{v} - \frac{<\text{vec}(\mathbf{B}_0), \mathbf{v}>}{f}\text{vec}(\mathbf{A}_0)\|_\infty$$
$$\leq 2\sqrt{2\psi_\tau}\mathbf{A}_0^{max}\|\mathbf{B}_0\|_2 p\sqrt{f}\sqrt{\frac{\log M}{n}} \tag{29}$$

Expanding $\tilde{\mathbf{A}}^1$:

$$\text{vec}(\tilde{\mathbf{A}}^1) = \frac{1}{f}\hat{\mathbf{R}}_A\text{vec}(\mathbf{Y}_1) - \text{vec}(\mathbf{A}_*)$$
$$= \frac{\text{tr}(\mathbf{B}_0(\mathbf{Y}_1 - \mathbf{Y}_*))}{f}\text{vec}(\mathbf{A}_0) + \text{vec}(\hat{\mathbf{A}}(\mathbf{B}_*) - \mathbf{A}_*)$$
$$+ \frac{1}{f}\tilde{\mathbf{R}}_A\text{vec}(\mathbf{Y}_1 - \mathbf{Y}_*) \tag{30}$$

where we used $\mathbf{R}_A = \text{vec}(\mathbf{A}_0)\text{vec}(\mathbf{B}_0^T)^T$ (see Eq. (91) from [4]).

Now, using the triangle inequality in (30), the bounds (27) and (29), the Cauchy-Schwarz inequality, we obtain w.h.p. (under conditions (24),(26),(28)), after some algebra:

$$\|\tilde{\mathbf{A}}^1\|_F \leq \sqrt{\frac{p}{f}}\|\mathbf{A}_0\|_2\|\mathbf{B}_0\|_2\|\mathbf{Y}_1 - \mathbf{Y}_*\|_F + p|\hat{\mathbf{A}}(\mathbf{B}_*) - \mathbf{A}_*|_\infty$$
$$+ \frac{1}{f}\|\tilde{\mathbf{R}}_A\|_2\|\mathbf{Y}_1 - \mathbf{Y}_*\|_F$$
$$\leq \tilde{C}_1(\sqrt{f} + pf^{-1/2})\sqrt{\frac{\log M}{n}} + \tilde{C}_2\sqrt{pf}\frac{\log M}{n} \tag{31}$$

where $\tilde{C}_1, \tilde{C}_2$ are absolute constants [12].

Let $c_1 > 0$. For

$$n \geq (\frac{\tilde{C}_2}{\tilde{C}_1 c_1})^2 p\log M \tag{32}$$

then, from (31), we have w.h.p.

$$\|\tilde{\mathbf{A}}^1\|_F \leq \tilde{C}_1(1+c_1)(\sqrt{f} + pf^{-1/2})\sqrt{\frac{\log M}{n}} \tag{33}$$

Using the permutation operator $\mathcal{R}(\cdot)$ defined in [4]:

$$\begin{aligned}\text{vec}(\mathcal{R}(\tilde{\mathbf{\Sigma}}_{FF}(2))) &= \text{vec}(\text{vec}(\tilde{\mathbf{A}}^1)\text{vec}(\mathbf{B}_*)^T) \\ &+ \text{vec}(\text{vec}(\mathbf{A}_*)\text{vec}(\tilde{\mathbf{B}}^0)^T) + \text{vec}(\text{vec}(\tilde{\mathbf{A}}^1)\text{vec}(\tilde{\mathbf{B}}^0)^T)\end{aligned} \tag{34}$$

From (25),(31), (34) and $\text{vec}(\tilde{\mathbf{\Sigma}}_{FF}(2)) = \mathbf{P}_R\text{vec}(\mathcal{R}(\tilde{\mathbf{\Sigma}}_{FF}(2)))$ [4], under conditions (24),(26),(28) and (32), w.h.p.,

$$\begin{aligned}\|\tilde{\mathbf{\Sigma}}_{FF}(2)\|_F &\leq \|\tilde{\mathbf{A}}_1\|_F\|\mathbf{B}_*\|_F \\ &+ \|\mathbf{A}_*\|_F\|\tilde{\mathbf{B}}^0\|_F + \|\tilde{\mathbf{A}}^1\|_F\|\tilde{\mathbf{B}}^0\|_F \\ &\leq \tilde{C}_3(p+2f)\sqrt{\frac{\log M}{n}} + \tilde{C}_4(f\sqrt{f/p}+\sqrt{pf})\frac{\log M}{n}\end{aligned} \tag{35}$$

where $\tilde{C}_3$ and $\tilde{C}_4$ are constants [12].

For

$$n \geq (\frac{\tilde{C}_4}{\tilde{C}_3 c_2})^2 \frac{(f\sqrt{f/p}+\sqrt{pf})^2}{(p+2f)^2}\log M$$

then, from (35) w.h.p.,

$$\|\tilde{\mathbf{\Sigma}}_{FF}(2)\|_F \leq \tilde{C}_3(1+c_2)(p+2f)\sqrt{\frac{\log M}{n}}$$

The proof for $k=2$ iterations is complete. Using a simple induction, it follows that the rate (13) holds for all k finite.

Next, we show that the convergence rate in the precision matrix Frobenius error is on the same order as the covariance matrix error. Let $\mathbf{\Theta}_{FF}(2) := \mathbf{\Sigma}_{FF}(2)^{-1}$. From (33), for

$$n > (\epsilon'\|\mathbf{X}_*\|_2\tilde{C}_1(1+c_1))^2(\sqrt{f}+pf^{-1/2})^2\log M$$

then w.h.p.,

$$\begin{aligned}\|\mathbf{X}_2-\mathbf{X}_*\|_F &\leq (1-\frac{1}{\epsilon'})^{-1}\|\mathbf{X}_*\|_2^2\tilde{C}_1(1+c_1) \\ &\times(\sqrt{f}+pf^{-1/2})\sqrt{\frac{\log M}{n}}\end{aligned} \tag{36}$$

Using (27) and (36), we have w.h.p.,

$$\begin{aligned}\|\mathbf{\Theta}_{FF}(2)-\mathbf{\Theta}_0\|_F &\leq \|\mathbf{X}_2-\mathbf{X}_*\|_F\|\mathbf{Y}_*\|_F \\ &+ \|\mathbf{Y}_1-\mathbf{Y}_*\|_F\|\mathbf{X}_*\|_F + \|\mathbf{X}_2-\mathbf{X}_*\|_F\|\mathbf{Y}_1-\mathbf{Y}_*\|_F \\ &\leq \tilde{D}_1(2f+p)\sqrt{\frac{\log M}{n}} + \tilde{D}_2(f\sqrt{\frac{f}{p}}+\sqrt{pf})\frac{\log M}{n}\end{aligned} \tag{37}$$

where $\tilde{D}_1$ and $\tilde{D}_2$ are constants.

For

$$n > (\frac{\tilde{D}_2}{\tilde{D}_1 d'})^2(\frac{f\sqrt{f/p}+\sqrt{pf}}{2f+p})^2\log M$$

the bound (37) becomes w.h.p.,

$$\|\mathbf{\Theta}_{FF}(2)-\mathbf{\Theta}_0\|_F \leq \tilde{D}_1(1+d')(2f+p)\sqrt{\frac{\log M}{n}}$$

Thus, the same rate $O_P\left(\sqrt{\frac{(p^2+f^2)\log M}{n}}\right)$ holds for the precision matrix Frobenius error.

■

APPENDIX E
PROOF OF THEOREM 3

Proof: We show that the first iteration of the KGL algorithm yields a fast statistical convergence rate of $O_P\left(\sqrt{\frac{(p+f)\log M}{n}}\right)$ by appropriately adjusting the regularization parameters. A simple induction finishes the proof. Adopt the notation from the proof of Thm. 2.

Lemma 2 implies that for

$$n \geq \frac{8(2+\tau)^2}{\psi_\tau}\log M \tag{38}$$

then w.h.p.,

$$|\tilde{\mathbf{B}}^0|_\infty \leq 2\sqrt{2\psi_\tau}\mathbf{B}_0^{max}\|\mathbf{A}_{init}^{-1}\mathbf{A}_0\|_2\sqrt{\frac{\log M}{np}} \tag{39}$$

where $\tilde{\mathbf{B}}^0 = \hat{\mathbf{B}}(\mathbf{A}_{init}) - \mathbf{B}_*$. From Proposition 1 and (39), we obtain w.h.p.,

$$\begin{aligned}\|\mathbf{Y}_1-\mathbf{Y}_*\|_F &\leq 2\sqrt{2}(1+c_y)\sqrt{1+c_{Y_0}}\|\mathbf{Y}_*\|_2^2 \\ &\times 2\sqrt{2\psi_\tau}\mathbf{B}_0^{max}\|\mathbf{A}_{init}^{-1}\mathbf{A}_0\|_2\sqrt{\frac{f\log M}{np}}\end{aligned} \tag{40}$$

where we also used $s_{Y_0} \leq c_{Y_0}f$ and $\mathbf{Y}_1 := \mathbf{G}(\hat{\mathbf{B}}(\mathbf{A}_{init}), \lambda_Y^{(1)}) = \mathbf{B}_1^{-1}$. Note that $fp^{-1}\log M = o(n)$ was used here.

Let $\grave{\mathbf{A}}^1 := \hat{\mathbf{A}}(\mathbf{B}_1) - \mathbf{A}_*$. Then, we have

$$\begin{aligned}\text{vec}(\grave{\mathbf{A}}^1) &= \frac{1}{f}\hat{\mathbf{R}}_A\text{vec}(\mathbf{Y}_1) - \text{vec}(\mathbf{A}_*) \\ &= \frac{\text{tr}(\mathbf{B}_0(\mathbf{Y}_1-\mathbf{Y}_*))}{f}\text{vec}(\mathbf{A}_0) + \text{vec}(\hat{\mathbf{A}}(\mathbf{B}_*)-\mathbf{A}_*) \\ &+ \frac{1}{f}\tilde{\mathbf{R}}_A\text{vec}(\mathbf{Y}_1-\mathbf{Y}_*)\end{aligned} \tag{41}$$

where we used $\mathbf{R}_A = \text{vec}(\mathbf{A}_0)\text{vec}(\mathbf{B}_0^T)^T$ (see Eq. (91) in [4]). Recall the definition of a mixed norm [16]:

$$\|\mathbf{W}\|_{\alpha,\beta} = \sup_{\|\mathbf{v}\|_\alpha=1}\|\mathbf{W}\mathbf{v}\|_\beta \tag{42}$$

From (40) and (42), we have w.h.p.,

$$\begin{aligned}\frac{1}{f}\|\tilde{\mathbf{R}}_A\text{vec}(\mathbf{Y}_1-\mathbf{Y}_*)\|_\infty &\leq \frac{\|\tilde{\mathbf{R}}_A\|_{2,\infty}}{f}\|\mathbf{Y}_1-\mathbf{Y}_*\|_F \\ &\leq 8\psi_\tau\mathbf{A}_0^{max}\mathbf{B}_0^{max}\|\mathbf{B}_0\|_2\|\mathbf{A}_{init}^{-1}\mathbf{A}_0\|_2 \\ &\times 2\sqrt{2}(1+c_y)\sqrt{1+c_{Y_0}}\|\mathbf{Y}_*\|^2\frac{1}{\sqrt{p}}\frac{\log M}{n}\end{aligned} \tag{43}$$

where we used the bound on the mixed-norm-i.e., for $n > \frac{8(2+\tau)^2}{\psi_\tau}\log M$, then w.h.p. (from Lemma 2),

$$\begin{aligned}\frac{\|\tilde{\mathbf{R}}_A\|_{2,\infty}}{f} &= \frac{1}{f}\sup_{\|\mathbf{v}\|_2=1}\|\tilde{\mathbf{R}}_A\mathbf{v}\|_\infty \\ &= \sup_{\|\mathbf{v}\|_2=1}\|\frac{1}{f}\hat{\mathbf{R}}_A\mathbf{v} - \frac{<\text{vec}(\mathbf{B}_0),\mathbf{v}>}{f}\text{vec}(\mathbf{A}_0)\|_\infty \\ &\leq 2\sqrt{2\psi_\tau}\mathbf{A}_0^{max}\|\mathbf{B}_0\|_2\sqrt{\frac{\log M}{nf}}\end{aligned}$$

From (41), applying the triangle inequality and using the Cauchy-Schwarz inequality, (40), (43), w.h.p.,

$$|\dot{\mathbf{A}}^1|_\infty \leq \frac{\sqrt{f}\|\mathbf{B}_0\|_2\|\mathbf{Y}_1-\mathbf{Y}_*\|_F}{f}|\mathbf{A}_0|_\infty + |\hat{\mathbf{A}}(\mathbf{B}_*)-\mathbf{A}_*|_\infty + \frac{1}{f}\|\tilde{\mathbf{R}}_A \text{vec}(\mathbf{Y}_1-\mathbf{Y}_*)\|_\infty \leq \bar{C}_1\left(\frac{1}{\sqrt{p}}+\frac{1}{\sqrt{f}}\right)\sqrt{\frac{\log M}{n}} + \bar{C}_2\frac{1}{\sqrt{p}}\frac{\log M}{n} \quad (44)$$

where we used Lemma 2 and $\bar{C}_1$ and $\bar{C}_2$ are absolute constants [12].

For

$$n > (\frac{\bar{C}_2}{\bar{C}_1 c'})^2 \frac{\log M}{(1+\sqrt{\frac{p}{f}})^2}$$

the bound in (44) simplifies to

$$|\dot{\mathbf{A}}^1|_\infty \leq \bar{C}_1(1+c')\left(\frac{1}{\sqrt{p}}+\frac{1}{\sqrt{f}}\right)\sqrt{\frac{\log M}{n}} \quad (45)$$

From Proposition 1 and (45), we obtain w.h.p.:

$$\|\mathbf{X}_1-\mathbf{X}_*\|_F \leq 2\sqrt{2}(1+c_x)\sqrt{1+c_{X_0}}\|\mathbf{X}_*\|_2^2\bar{C}_1(1+c') \times\left(1+\sqrt{\frac{p}{f}}\right)\sqrt{\frac{\log M}{n}} \quad (46)$$

where we used $s_{X_0} \leq c_{X_0}p$ and $\mathbf{X}_1 := \mathbf{G}(\hat{\mathbf{A}}(\mathbf{B}_1), \lambda_X^{(1)})$, $\mathbf{X}_* := \mathbf{A}_*^{-1}$. Note that $(1+\sqrt{p/f})^2\log M = o(n)$ was used here.

Finally, using (40) and (46), we obtain w.h.p.:

$$\|\mathbf{\Theta}_{KGL}(2)-\mathbf{\Theta}_0\|_F = \|\mathbf{X}_1\otimes\mathbf{Y}_1-\mathbf{X}_*\otimes\mathbf{Y}_*\|_F \leq \|\mathbf{Y}_1-\mathbf{Y}_*\|_F\sqrt{p}\|\mathbf{X}_*\|_2 + \|\mathbf{X}_1-\mathbf{X}_*\|_F\sqrt{f}\|\mathbf{Y}_*\|_2 + \|\mathbf{Y}_1-\mathbf{Y}_*\|_F\|\mathbf{X}_1-\mathbf{X}_*\|_F \leq \bar{C}_3(2\sqrt{f}+\sqrt{p})\sqrt{\frac{\log M}{n}} + \bar{C}_4(1+\sqrt{\frac{f}{p}})\frac{\log M}{n} \quad (47)$$

where $\bar{C}_3$ and $\bar{C}_4$ are constants [12]. For

$$n > (\frac{\bar{C}_4}{\bar{C}_3 c''})^2\left(\frac{1+\sqrt{f/p}}{2\sqrt{f}+\sqrt{p}}\right)^2\log M$$

the bound (47) further becomes:

$$\|\mathbf{\Theta}_{KGL}(2)-\mathbf{\Theta}_0\|_F \leq \bar{C}_3(1+c'')(2\sqrt{f}+\sqrt{p})\sqrt{\frac{\log M}{n}}$$

Note that $\|\mathbf{\Theta}_{KGL}(2)-\mathbf{\Theta}_0\|_F^2 = O_P\left(\frac{(p+f+\sqrt{pf})\log M}{n}\right) = O_P\left(\frac{(p+f)\log M}{n}\right)$ as $p, f, n \to \infty$. This concludes the first part of the proof. The rest of the proof follows by similar bounding arguments coupled with induction. The rate remains the same as the number of iterations increases, but the constant on front may change.

Next, we show that the convergence rate in the covariance matrix Frobenius error is on the same order as the inverse. From (40), for

$$n > (\epsilon' 2\sqrt{2}(1+c_y)\sqrt{1+c_{Y_0}})^2\|\mathbf{Y}_*\|_2^2\kappa(\mathbf{Y}_0)^2 f p^{-1}\log M$$

we have w.h.p. $\lambda_{min}(\mathbf{Y}_1) \geq \lambda_{min}(\mathbf{Y}_*) - \|\mathbf{Y}_1-\mathbf{Y}_*\|_F \geq (1-\frac{1}{\epsilon'})\lambda_{min}(\mathbf{Y}_*)$, which in turn implies w.h.p.,

$$\|\mathbf{B}_1-\mathbf{B}_*\|_F \leq \left(1-\frac{1}{\epsilon'}\right)^{-1} 2\sqrt{2}(1+c_y)\sqrt{1+c_{Y_0}}\kappa(\mathbf{Y}_0)^2 \times\sqrt{\frac{f}{p}}\sqrt{\frac{\log M}{n}} \quad (48)$$

[6] Using a similar argument, from (46), for

$$n > \frac{(\epsilon' 2\sqrt{2}(1+c_x)\sqrt{1+c_{X_0}}\bar{C}_1(1+c'))^2}{\lambda_{min}(\mathbf{A}_*)^2}\kappa(\mathbf{X}_0)^2 \times(1+\sqrt{\frac{p}{f}})^2\log M$$

we have w.h.p.,

$$\|\mathbf{A}_1-\mathbf{A}_*\|_F \leq \left(1-\frac{1}{\epsilon'}\right)^{-1} 2\sqrt{2}(1+c_x)\sqrt{1+c_{X_0}}\bar{C}_1(1+c') \times\kappa(\mathbf{X}_0)^2\left(1+\sqrt{\frac{p}{f}}\right)\sqrt{\frac{\log M}{n}} \quad (49)$$

where $\mathbf{A}_1 = \mathbf{X}_1^{-1}$.

Let $\mathbf{\Sigma}_{KGL}(2) := \mathbf{\Theta}_{KGL}(2)^{-1} = \mathbf{A}_1\otimes\mathbf{B}_1$. Then, w.h.p.,

$$\|\mathbf{\Sigma}_{KGL}(2)-\mathbf{\Sigma}_0\|_F \leq \|\mathbf{A}_1-\mathbf{A}_*\|_F\|\mathbf{B}_*\|_F + \|\mathbf{B}_1-\mathbf{B}_*\|_F\|\mathbf{A}_*\|_F + \|\mathbf{A}_1-\mathbf{A}_*\|_F\|\mathbf{B}_1-\mathbf{B}_*\|_F \leq \bar{D}_1(2\sqrt{f}+\sqrt{p})\sqrt{\frac{\log M}{n}} + \bar{D}_2(1+\sqrt{\frac{f}{p}})\frac{\log M}{n} \quad (50)$$

where $\bar{D}_1$ and $\bar{D}_2$ are constants [12]. For

$$n > (\frac{\bar{D}_2}{\bar{D}_1 d})^2\left(\frac{1+\sqrt{\frac{f}{p}}}{2\sqrt{f}+\sqrt{p}}\right)^2\log M$$

then (50) implies w.h.p.,

$$\|\mathbf{\Sigma}_{KGL}(2)-\mathbf{\Sigma}_0\|_F \leq \bar{D}_1(1+d)(2\sqrt{f}+\sqrt{p})\sqrt{\frac{\log M}{n}}$$

Thus, the same rate $O_P\left(\sqrt{\frac{(p+f)\log M}{n}}\right)$ holds for the error in the covariance matrix.

■

REFERENCES

[1] 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.

[2] M. Yuan and Y. Lin, "Model selection and estimation in the gaussian graphical model." *Biometrika*, vol. 94, pp. 19–35, 2007.

[3] O. Banerjee, L. E. 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, pp. 485–516, March 2008.

[4] K. Werner, M. Jansson, and P. Stoica, "On estimation of covariance matrices with Kronecker product structure," *IEEE Transactions on Signal Processing*, vol. 56, no. 2, February 2008.

[5] K. Werner and M. Jansson, "Estimation of kronecker structured channel covariances using training data," in *Proceedings of EUSIPCO*, 2007.

[6] Here, $\mathbf{B}_1 = \mathbf{Y}_1^{-1}$ exists since $\mathbf{Y}_1$ is positive definite (see (10)).

[6] N. Lu and D. Zimmerman, "On likelihood-based inference for a separable covariance matrix," Statistics and Actuarial Science Dept., Univ. of Iowa, Iowa City, IA, Tech. Rep., 2004.
[7] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.
[8] S. Zhou, J. Lafferty, and L. Wasserman, "Time varying undirected graphs," *Journal of Machine Learning Research*, vol. 80, pp. 295–319, 2010.
[9] A. Rothman, P. Bickel, E. Levina, and J. Zhu, "Sparse permutation invariant covariance estimation," *Electronic Journal of Statistics*, vol. 2, pp. 494–515, 2008.
[10] P. Ravikumar, M. Wainwright, G. Raskutti, and B. Yu, "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence," *Advances in Neural Information Processing Systems*, 2008.
[11] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
[12] T. Tsiligkaridis, A. Hero, and S. Zhou, "Convergence properties of kronecker graphical lasso algorithms," *arXiv:submit/0448399*, April 2012.
[13] C.-J. Hsieh, M. A. Sustik, I. S. Dhillon, and P. Ravikumar, "Sparse inverse covariance matrix estimation using quadratic approximation," *Advances in Neural Information Processing Systems*, vol. 24, 2011.
[14] R. G. Bartle and D. R. Sherbert, *Introduction to Real Analysis*. John Wiley & Sons, 2000.
[15] N. Lu and D. Zimmerman, "The likelihood ratio test for a separable covariance matrix," *Statistics and Probability Letters*, vol. 73, no. 5, pp. 449–457, May 2005.
[16] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1990.