Sharp rates of convergence for the tensor graphical Lasso estimator

Kristjan Greenewald
Kristjan.H.Greenewald@ibm.com
MIT-IBM Watson AI Lab,
Cambridge, MA 02141

Shuheng Zhou
szhou@ucr.edu
Department of Statistics
University of California, Riverside
Riverside, CA 92521

Abstract

Many modern datasets exhibit dependencies among observations as well as variables. This gives rise to the challenging problem of analyzing high-dimensional matrix-variate data with unknown dependence structures. To address this challenge, Kalaitzis et. al. (2013) proposed the Bigraphical Lasso (BiGLasso), an estimator for precision matrices of matrix-normals based on the Cartesian product of graphs. Subsequently, Greenewald, Zhou and Hero (GZH 2019) introduced a multiway tensor generalization of the BiGLasso estimator, known as the TeraLasso estimator. Building upon GZH 2019, we provide sharper rates of convergence in the Frobenius and operator norm for both BiGLasso and TeraLasso estimators for estimating inverse covariance matrices. In particular, (a) we strengthen the bounds for the relative errors in the operator and Frobenius norm by a factor of approximately $\log p$; (b) Crucially, this improvement allows for finite-sample estimation errors in both norms to be derived for the two-way Kronecker sum model. The two-way regime is particularly significant since it is the setting of common and generic applications in practice. Normality is not needed in our analysis; instead, we consider subgaussian ensembles and derive tight concentration of measure bounds, using tensor unfolding techniques. The proof techniques may be of independent interest to the analysis of tensor-valued data.

Keywords: Subgaussian concentration, Tensor data, graphical models, Kronecker sum, matrix-variate normal

1. Introduction

Matrix and tensor-valued data are ubiquitous in modern statistics and machine learning, flowing from sources as diverse as medical and radar imaging modalities, spatial-temporal and meteorological data collected from sensor networks and weather stations, and biological, neuroscience and spatial gene expression data aggregated over trials and time points. Learning useful structures from these large scale, high-dimensional data with complex dependencies in the low sample regime is an important task. Undirected graphs are often used to describe high dimensional distributions. Under sparsity conditions, the graph can be estimated using ℓ_1-penalization methods, such as the graphical Lasso (GLasso) Friedman et al. (2008) and multiple (nodewise) regressions Meinshausen and Bühlmann (2006). Under suitable

conditions, including the independence among samples, such approaches yield consistent (and sparse) estimation in terms of graphical structure and fast convergence rates with respect to the operator and Frobenius norm for the covariance matrix and its inverse. The independence assumptions as mentioned above substantially simplify mathematical derivations but they tend to be very restrictive.

To remedy this, recent work has demonstrated another regime where further improvements in the sample size lower bounds are possible under additional structural assumptions, which arise naturally in the above mentioned contexts for data with complex dependencies. For example, the matrix-normal model Dawid (1981) as studied in Allen and Tibshirani (2010); Leng and Tang (2012); Tsiligkaridis et al. (2013); Zhou (2014) restricts the topology of the graph to tensor product graphs where the precision matrix corresponds to a Kronecker product over two component graphs. Moreover, Zhou (2014) showed that one can estimate the covariance and inverse covariance matrices well using only one instance from the matrix variate normal distribution. However, such a normality assumption is also not needed, as elaborated in a recent paper by the same author Zhou (2023). More specifically, while the precision matrix encodes conditional independence relations for Gaussian distributions, for the more general subgaussian matrix variate model, this no longer holds. However, the inverse covariance matrix still encodes certain zero correlation relations between the residual errors and the covariates in a regression model, analogous to the commonly known Gaussian regression model Lauritzen (1996). See Zhou (2023), where such regression model is introduced for subgaussian matrix variate data. See also Gupta and Varga (1992); Hornstein et al. (2019) and references therein for more recent applications of matrix variate models.

Along similar lines, the Bigraphical Lasso framework was proposed to parsimoniously model conditional dependence relationships of matrix variate data based on the Cartesian product of graphs. As pointed out in Kalaitzis et al. (2013), the associativity of the Kronecker sum yields an approach to the modeling of datasets organized into 3 or higher-order tensors. In Greenewald et al. (2019a), we explored this possibility to a great extent, by (a) introducing the tensor graphical Lasso (TeraLasso) procedure for estimating sparse K-way decomposable inverse covariance matrices for all $K \geq 2$; and (b) showing the rates of convergence in the Frobenius and the operator norm for estimating this class of inverse covariance matrices for subgaussian tensor-valued data. When the data indeed follows a matrix normal model, TeraLasso also effectively recovers the conditional dependence graphs and precision matrices simultaneously for a class of Gaussian graphical models by restricting the topology to Cartesian product graphs. We provided a composite gradient-based optimization algorithm, and obtained algorithmic and statistical rates of convergence for estimating structured precision matrix for tensor-valued data.

To make the subsequent discussion tangible, we introduce the following notation. Consider the mean zero K-order random tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$, and assume that we are given n independent samples $\boldsymbol{\mathcal{X}}_1, \ldots, \boldsymbol{\mathcal{X}}_n \sim \boldsymbol{\mathcal{X}}$. Here $\sim$ represents that two vectors follow the same distribution. Denote by $\mathbf{p} = [d_1, \ldots, d_K]$ the vector of component dimensions, p the product of d_js. Hence

$$\text{vec}(\boldsymbol{\mathcal{X}}) \in \mathbb{R}^p, \quad \text{where} \quad p = \prod\nolimits_k d_k \quad \text{and} \quad m_k = \prod\nolimits_{i \neq k} d_i = p/d_k$$

is the effective sample size we have to estimate the relations among the d_k features in the matrix variate models. It was shown that due to the element replication inherent in the Cartesian product structure, the precision matrix in the TeraLasso model can be accurately estimated from limited data samples of high dimensional variables with multiway coordinates such as space, time and replicates. See Greenewald et al. (2019a) and Lemma 1 in the present work for elaborations. In particular, single sample convergence was proved for $K > 2$ and empirically observed for all K. In contrast, direct application of the model in Friedman et al. (2008) and the analysis framework in Rothman et al. (2008); Zhou et al. (2011) typically require the sample size n to scale as proportionally to p, at least when the total number of edges is $\propto p$, which is still often too large to be practical. As a result, it is common to assume certain axes of $\mathcal{X}$ are i.i.d., often an overly simplistic model.

Contributions. In the present work, we improve on the statistical convergence rates provided in Greenewald et al. (2019a), using tighter concentration bounds on the diagonal component of the trace term in the loss function; cf. (6). We strengthen the bounds for the relative errors in the operator and Frobenius norm by a factor of approximately $\log p$. Crucially, this improvement allows for finite-sample Frobenius and operator norm estimation guarantees to be extended from $K > 2$ to the two-way Kronecker sum model. This closes the gap between the low single-sample error for the two-way model observed in Greenewald et al. (2019a) and the theoretical bounds presented therein. The main theoretical results are stated in Theorem 4 and its corollaries. In general, K-way models can be taken as refinements of the Bigraphical models, due to the associativity of the Kronecker sum. Finally, we provide experiments that confirm the derived rates and demonstrate accurate estimation in the single-sample and replicated-sample regimes.

The key innovation in our convergence analysis is the uniform concentration of measure bounds on diagonal components of the K trace terms in the loss function (9), to be stated in Lemma 2. Lemma 2 combines new proof ideas based on sharp covering net arguments and the orthogonal decomposition Lemma 21, as already established in our previous paper Greenewald et al. (2019a). In the present work, due to the tighter error bound on the diagonal component of the loss function as stated in Lemma 2, we achieve the sharper rates of convergence in Theorem 4, which significantly improve upon Theorem 5 as originally proved in Greenewald et al. (2019a). Specifically, we replace the $p \log p$ factor with p for the relative errors in the operator and Frobenius norm in Theorem 4, thanks to the improved analysis on the diagonal component.

1.1 Related Work

Models similar to the Kronecker sum precision model have been successfully used in a variety of fields, including regularization of multivariate splines Wood (2006); Eilers and Marx (2003); Kotzagiannidis and Dragotti (2017); Wood et al. (2016), design of physical networks Imrich et al. (2008); Van Loan (2000); Fey et al. (2018), and Sylvester equations arising from the discretization of separable K-dimensional PDEs with tensorized finite elements Grasedyck (2004); Kressner and Tobler (2010); Beckermann et al. (2013); Shi et al. (2013); Ellner (1986). Additionally, Kronecker sums find extensive use in applied mathematics and statistics, including beam propagation physics Andrianov (1997), control theory Luenberger (1966); Chapman et al. (2014), fluid dynamics Dorr (1970), errors-in-

variables Rudelson and Zhou (2017), and spatio-temporal neural processes Schmitt et al. (2001). Additional discussions and motivations for the model to be considered in applied settings can be found in Kalaitzis et al. (2013); Greenewald et al. (2019a); Li et al. (2022). Subsequent to Greenewald et al. (2019a), Wang and Hero (2021) presented SG-PALM, a related model where the precision matrix is the *square* of a K-way Kronecker sum. This model is motivated by a connection to solutions of random Sylvester equations, and has been applied to challenging problems in climate prediction. In contrast to TeraLasso, the statistical convergence bounds presented for SG-PALM do not imply single-sample convergence and instead require a number of samples proportional to the dimension; See also Wang et al. (2022) for a recent survey.

Recently, several methods have arisen that can speed up the numerical convergence of the optimization of the BiGLasso objective of Kalaitzis et al. (2013), which is equivalent to the $K = 2$ TeraLasso objective. A Newton-based optimization algorithm for $K = 2$ was presented in Yoon and Kim (2022) that provides significantly faster convergence in ill-conditioned settings. Subsequently, Li et al. (2022) also developed a scalable flip-flop approach, building upon the original BiGLasso flip-flop algorithm as derived in Kalaitzis et al. (2013). Using the Kronecker sum eigenvalue decomposition similar to that of Greenewald et al. (2019a) to make the memory requirements scalable, their algorithm also provides faster numerical convergence than the first-order algorithm presented in Greenewald et al. (2019a), especially when the precision matrix is ill-conditioned. They also provided a Gaussian copula approach for applying the model to certain non-Gaussian data. As our focus in the current work is the statistical convergence rate, not optimization convergence, we continue to use the first-order optimization algorithm of Greenewald et al. (2019a) in our experiments because it is flexible to $K > 2$. For $K = 2$, it will recover the same estimate as the algorithms of Li et al. (2022) and Yoon and Kim (2022) since the objective function is convex and all three algorithms are guaranteed to converge to the unique global minimum. Although non-convex regularizers were also considered in Greenewald et al. (2019a), we do not consider these in the present work. Instead, we focus on deriving sharp statistical rates of convergence in the Frobenius and operator norm for estimating precision matrix (4) with convex function (6).

Organization. The rest of the paper is organized as follows. In Section 2, we define our model and the method, as well as discussions on our method. Section 3 presents our main technical results, namely, Theorem 4 and its corollaries, and discussions on our results. In Section 4, we elaborate upon the proof strategies for Theorems 4 and 5. Section 5 presents key technical proofs regarding the diagonal component of the loss function, highlighting tensor unfolding techniques and Hanson-Wright inequalities, which are crucial in analyzing the subgaussian tensor-valued data with complex dependencies. In Section 6, we show numerical results that validate our theoretical predictions. We conclude in Section 7. Additional technical proofs appear in the appendix.

1.2 Definitions and notations

Let $e_1, \ldots, e_n$ be the canonical basis of $\mathbb{R}^n$. Let B_2^n and $\mathbb{S}^{n-1}$ be the unit Euclidean ball and the unit sphere of $\mathbb{R}^n$, respectively. For a set $J \subset \{1, \ldots, n\}$, denote $E_J = \text{span}\{e_j : j \in J\}$. We denote by $[n]$ the set $\{1, \ldots, n\}$. We refer to a vector $v \in \mathbb{R}^n$ with at most $d \in [n]$

nonzero entries as a d-sparse vector. For a finite set V, the cardinality is denoted by $|V|$. For a vector $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, denote by $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$ and $|x|_1 := \sum_j |x_j|$. For a given vector $x \in \mathbb{R}^m$, $\mathrm{diag}(x)$ denotes the diagonal matrix whose main diagonal entries are the entries of x.

We use A for matrices, $\boldsymbol{\mathcal{A}}$ for tensors, and $\mathbf{a}$ for vectors. For $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{d_1 \times d_2 \ldots \times d_N}$, we use $\mathrm{vec}(\boldsymbol{\mathcal{A}}) \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_N}$ as in Kolda and Bader (2009), and define $\boldsymbol{\mathcal{A}}^T \in \mathbb{R}^{d_N \times \ldots d_2 \times d_1}$ by analogy to the matrix transpose, i.e. $[\boldsymbol{\mathcal{A}}^T]_{i_1,\ldots,i_N} = \boldsymbol{\mathcal{A}}_{i_N,\ldots,i_1}$. For a symmetric matrix A, let $\phi_{\max}(A)$ and $\phi_{\min}(A)$ be the largest and the smallest eigenvalue of A respectively. For a matrix A, we use $\|A\|_2$ to denote its operator norm and $\|A\|_F$ the Frobenius norm, given by $\|A\|_F = (\sum_{i,j} a_{ij}^2)^{1/2}$. For a matrix $A = (a_{ij})$ of size $m \times n$, let $\|A\|_\infty = \max_i \sum_{j=1}^n |a_{ij}|$ and $\|A\|_1 = \max_j \sum_{i=1}^m |a_{ij}|$ denote the maximum absolute row and column sum of the matrix A respectively. Let $\|A\|_{\max} = \max_{i,j} |a_{ij}|$ denote the componentwise matrix maximum norm. Let $\mathrm{diag}(A)$ be the diagonal of A. Let $\mathrm{offd}(A) = A - \mathrm{diag}(A)$. Let $\kappa(A) = \phi_{\max}(A)/\phi_{\min}(A)$ denote the condition number for matrix A. Denote by $|A|$ the determinant of A. We use the inner product $\langle A, B \rangle = \mathrm{tr}(A^T B)$ throughout this paper. The inner product of two same-sized tensors $\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}} \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_N}$ is sum of the products of their entries, i.e.,

$$\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}} \rangle = \sum_{i_1=1}^{d_1} \sum_{i_2=1}^{d_2} \cdots \sum_{i_N=1}^{d_N} x_{i_1 i_2 \ldots\ldots i_N} y_{i_1 i_2 \ldots i_N} \tag{1}$$

where $x_{i_1,\ldots,i_N}$ denotes the $(i_1, \ldots, i_N)$ element of $\boldsymbol{\mathcal{X}}$. Fibers are the higher-order analogue of matrix rows and columns. Following definition by Kolda and Bader (2009), tensor unfolding or matricization of $\boldsymbol{\mathcal{X}}$ along the kth-mode is denoted as $\mathbf{X}^{(k)}$, and is formed by arranging the mode-k fibers as columns of the resulting matrix of dimension $d_k \times m_k$. Denote by $\mathbf{X}^{(k)T}$ its transpose. In general, when we have K way tensor, matrix $\mathbf{X}^{(k)}$ of dimension $d_k \times m_k$, is formed by rearranging the mode-k fibers of a tensor $\boldsymbol{\mathcal{X}}$ as columns, a process commonly referred to as tensor unfolding or flattening of $\boldsymbol{\mathcal{X}}$ along the k^{th} mode.

When extracted from the tensor, fibers are always assumed to be oriented as column vectors. One can compute the corresponding (rescaled) Gram matrix S^k with $\mathbf{X}^{(k)}\mathbf{X}^{(k)T}/m_k$. Denote by $X_j^{(k)}$ the j^{th} column vector of $\mathbf{X}^{(k)} \in \mathbb{R}^{d_k \times m_k}$. Thus, we have m_k columns to compute $m_k S^k$,

$$m_k S^k = \mathbf{X}^{(k)}\mathbf{X}^{(k)T} = \sum_{j=1}^{m_k} X_j^{(k)} \otimes X_j^{(k)} = \tag{2}$$
$$\sum_{i_1=1}^{d_1} \cdots \sum_{i_{k-1}=1}^{d_{k-1}} \sum_{i_{k+1}=1}^{d_{k+1}} \cdots \sum_{i_K=1}^{d_K} \mathbf{x}_{i_1 \ldots i_{k-1} : i_{k+1} \ldots i_K} \otimes \mathbf{x}_{i_1 \ldots i_{k-1} : i_{k+1} \ldots i_K},$$

where $\mathbf{x}_{i_1 \ldots i_{k-1} : i_{k+1} \ldots i_K}$ denotes a mode-k fiber of tensor $\boldsymbol{\mathcal{X}}$ with all indices except for one being fixed, and the tensor product for a vector x denotes $x \otimes x = xx^T$. As mentioned in Kolda and Bader (2009), the specific permutation of columns is not important so long as it is consistent across related calculations. The vectorization of a K-way tensor $\boldsymbol{\mathcal{X}}$ is denoted as $\mathrm{vec}(\boldsymbol{\mathcal{X}})$ and is defined in the standard way as in Kolda and Bader (2009), by analogy to the vectorization of matrices. As an example, suppose we fix $K = 3$. Third-order tensors $\boldsymbol{\mathcal{X}}$

have column, row and tube fibers denoted by $\mathbf{x}_{:jk}, \mathbf{x}_{i:k}$, and $\mathbf{x}_{ij:}$, respectively. Slices are two-dimensional sections of a tensor, defined by fixing all but two indices, denote by $\mathbf{X}_{i::}$, $\mathbf{X}_{:j:}$ and $\mathbf{X}_{::k}$ respectively. For two numbers a, b, $a \wedge b := \min(a, b)$, and $a \vee b := \max(a, b)$. We write $a \asymp b$ if $ca \leq b \leq Ca$ for some positive absolute constants c, C that are independent of n, m, sparsity, and sampling parameters. We write $f = O(g)$ or $f \ll g$ if $|f| \leq Cg$ for some absolute constant $C < \infty$ and $f = \Omega(g)$ or $f \gg g$ if $g = O(f)$. We write $f = o(g)$ if $f/g \to 0$ as $n \to \infty$, where the parameter n will be the size of the matrix under consideration. In this paper, $C, c, c', C_1, C_2, \ldots$, etc, denote various absolute positive constants which may change line by line.

2. The model and the method

For a random variable Z, the subgaussian (or ψ_2) norm of Z denoted by $\|Z\|_{\psi_2}$ is defined as $\|Z\|_{\psi_2} = \inf\{t > 0 \,:\, \mathbb{E}\exp(Z^2/t^2) \leq 2\}$. Consider the tensor-valued data $\mathfrak{X}$ generated from a subgaussian random vector $Z = (Z_j) \in \mathbb{R}^p$ with independent mean-zero unit variance components whose ψ_2 norms are uniformly bounded:

$$\mathrm{vec}\{\,\mathfrak{X}\,\} \;=\; \Sigma_0^{1/2} Z, \quad \text{where } \mathbb{E}\,(Z_j) = 0,\ \mathbb{E}Z_j^2 = 1, \ \text{ and } \|Z_j\|_{\psi_2} \leq C_0, \forall i, j. \tag{3}$$

We refer to $\mathfrak{X} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ as an order-K subgaussian random tensor with covariance $\Sigma_0 \in \mathbb{R}^{p \times p}$ throughout this paper. To simplify the multiway Kronecker notation, we define

$$I_{[d_k:\ell]} = \underbrace{I_{d_k} \otimes \cdots \otimes I_{d_\ell}}_{\ell - k + 1 \text{ factors}}$$

where $\otimes$ denotes the Kronecker (direct) product and $\ell \geq k$. We assume that the precision matrix Ω_0 (inverse of covariance Σ_0) of $\mathfrak{X}$ is the K-way Kronecker sum of matrix components $\{\Psi_k\}_{k=1}^K$:

$$\Omega_0 := \Sigma_0^{-1} = \Psi_1 \oplus \cdots \oplus \Psi_K = \sum_{k=1}^{K} I_{[d_{1:k-1}]} \otimes \Psi_k \otimes I_{[d_{k+1:K}]}. \tag{4}$$

We denote the set of possible precision matrices expressible in this form as $\mathcal{K}_{\mathbf{p}}^{\sharp}$. Specifically, $\mathcal{K}_{\mathbf{p}}^{\sharp}$ is the set of positive definite matrices that are decomposable into a Kronecker sum of fixed factor dimensions $\mathbf{p} = [d_1, \ldots, d_K]$:

$$\mathcal{K}_{\mathbf{p}}^{\sharp} = \{A \succ 0 | A \in \mathcal{K}_{\mathbf{p}}\},\ \ \mathcal{K}_{\mathbf{p}} = \{A \in \mathbb{R}^{p \times p} : \exists\, B_k \in \mathbb{R}^{d_k \times d_k} \text{ s.t. } A = B_1 \oplus \cdots \oplus B_K\}. \tag{5}$$

The TeraLasso estimator minimizes the negative ℓ_1-penalized Gaussian log-likelihood function over the domain $\mathcal{K}_{\mathbf{p}}^{\sharp}$ of precision matrices Ω having Kronecker sum form Greenewald et al. (2019a):

$$\widehat{\Omega} \;=\; \arg\min_{\Omega \in \mathcal{K}_{\mathbf{p}}^{\sharp}} \left\{ -\log|\Omega| + \,\langle\, \widehat{S}, \Omega \,\rangle\, + \sum_{k=1}^{K} m_k \rho_{n,k}\, |\Psi_k|_{1,\mathrm{off}} \right\}, \tag{6}$$

$$\text{where} \qquad \widehat{S} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{vec}\{\,\mathfrak{X}_i^{\mathrm{T}}\,\}(\mathrm{vec}\{\,\mathfrak{X}_i^{\mathrm{T}}\,\})^{\mathrm{T}} \quad \text{and} \quad |\Psi_{\mathrm{k}}|_{1,\mathrm{off}} = \sum_{\mathrm{i} \neq \mathrm{j}} |\Psi|_{\mathrm{k,ij}}. \tag{7}$$

Here the $\mathfrak{X}_i \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ are n i.i.d. samples from the data distribution. Throughout this paper, we refer to n as the number of samples for the tensor data. Lemma 9 of Greenewald et al. (2019b) reveals that the $p \times p$ sample outer product $\text{vec}\{\mathfrak{X}^{\mathrm{T}}\}\text{vec}\{\mathfrak{X}^{\mathrm{T}}\}^{\mathrm{T}}$ only enters into the objective function (6) through its lower dimensional projections $S^k \in \mathbb{R}^{d_k \times d_k}$ (2). Thus, the mode-k Gram matrix S_n^k and its corresponding factor-wise marginal covariance $\Sigma_0^{(k)} = \mathbb{E}[S^k]$ are given by

$$S_n^k = \frac{1}{nm_k} \sum_{i=1}^{n} \mathbf{X}^{(k,i)} [\mathbf{X}^{(k,i)}]^T \quad \text{and} \quad \Sigma_0^{(k)} = \frac{1}{m_k} \mathbb{E}[\mathbf{X}^{(k)} \mathbf{X}^{(k)T}], \quad k = 1, \ldots, K, \tag{8}$$

respectively. Here $\mathbf{X}^{(k,i)}, i \in [n]$ are i.i.d. copies of $\mathbf{X}^{(k)}, \forall k$ as in (2), derived from independent tensors $\mathfrak{X}_i$. Then for $\mathcal{K}_{\mathbf{p}}^{\sharp}$, the set of positive definite Kronecker sum matrices, we have [1]

$$\widehat{\Omega} \;=\; \underset{\Omega \in \mathcal{K}_{\mathbf{p}}^{\sharp}}{\arg\min}\, Q(\Omega) := \underset{\Omega \in \mathcal{K}_{\mathbf{p}}^{\sharp}}{\arg\min} \left\{ -\log|\Omega| + \sum_{k=1}^{K} m_k \left(\langle S_n^k, \Psi_k \rangle + \rho_{n,k} \, |\Psi_k|_{1,\text{off}} \right) \right\}. \tag{9}$$

Throughout this paper, the subscript n is omitted from S_n^k and $\rho_{n,k}$ ($\delta_{n,k}$) in case $n = 1$ to avoid clutter in the notation.

Discussions. Clearly, we have replaced the trace term $\langle \widehat{S}, \Omega \rangle$ in (6) with the weighted sum over component-wise trace terms in (9), in view of (4); cf. Lemma 1. Here, the weight $m_k = p/d_k$ for each k is determined by the number of times for which a structure Ψ_k is replicated in Ω_0. Correspondingly, the following events $\{\mathcal{T}_k, k = 1, \ldots, K\}$ as originally defined in Greenewald et al. (2019a) and used in the present work reflect this aggregation with nm_k being the effective size for estimating Ψ_k. We define event $\mathcal{T} = \bigcap_{i=1}^{K} \mathcal{T}_k$, where

$$\text{event} \;\; \mathcal{T}_k \;\; = \;\; \left\{ \max_{i \neq j} \left| S_{n,ij}^k - \Sigma_{0,ij}^{(k)} \right| \leq \delta_{n,k}, \quad \text{for} \quad \delta_{n,k} \asymp \|\Sigma_0\|_2 \sqrt{\frac{\log p}{nm_k}} \right\}; \tag{10}$$

cf. (A1). Event $\mathcal{T}$ is needed to control the off-diagonal component of the loss function as in Lemma 10 (cf. Proof of Lemma 12 Greenewald et al. (2019b)). Here and in Greenewald et al. (2019a), the penalty parameters $\rho_{n,k}$ are chosen to dominate the maximum of entrywise errors for estimating the population $\Sigma_0^{(k)}$ with sample S_n^k as in (8) and (2) for each k on event $\mathcal{T}$. This choice works equally well for the subgaussian model (3). Intuitively, we use nm_k fibers to estimate relations between and among the d_k features along the k^{th} mode as encoded in Ψ_k and this allows optimal statistical rates of convergence to be derived, in terms of entrywise errors for estimating $\Sigma_0^{(k)}$ with S_n^k. As we will show in Theorem 5 Greenewald et al. (2019a), these entrywise error bounds already enabled a significant improvement in the sample size lower bound in order to estimate parameters and the associated conditional dependence graphs along different modes.

However, they are not sufficient to achieve operator norm-type of bounds as we elaborate in Theorem 4 for inverse covariance estimation. In fact, using entrywise bounds to control

1. We usually use this form (9) of the objective function, which depends on the training data via the coordinate-wise Gram matrices S_n^k. Besides making the structure more clear, this expression is more friendly to computation as it involves $d_k \times d_k$ component matrices rather than $p \times p$ matrices.

the diagonal components of the trace terms results in an extra $\log p$ factor in the sample size lower bound and correspondingly a slower rate of convergence. This extraneous $\log p$ factor is undesirable since the diagonal component of the loss function dominates the overall rate of convergence in sparse settings for inverse covariance estimation; cf. cf. (A3) and Theorems 5 and 4, and discussions that immediately follow.

For $K = 2$ and $\Omega_0 = \Psi_1 \oplus \Psi_2 = \Psi_1 \otimes I_{d_1} + I_{d_2} \otimes \Psi_2$, the objective function (9) is similar in spirit to the BiGLasso objective of Kalaitzis et al. (2013), where $S_n^k, k = 1, 2$ correspond to the Gram matrices computed from row and column vectors of matrix variate samples $X_1, \ldots, X_n \in \mathbb{R}^{d_1 \times d_2}$ respectively. When $\Omega_0 = \Psi_1 \otimes \Psi_2$ is a Kronecker product rather than a Kronecker sum over the factors, the objective function (9) is also closely related to estimators in Zhou (2014), where $\log |\Omega_0|$ is a linear combination of $\log |\Psi_k|, k = 1, 2$. When $\mathfrak{X}$ follows a multivariate Gaussian distribution and the precision matrix Ω_0 has a decomposition of the form (4), the sparsity pattern of Ψ_k for each k corresponds to the conditional independence graph across the k^{th} dimension of the data; See Kalaitzis et al. (2013); Greenewald et al. (2019a). Similar to the graphical Lasso, incorporating an ℓ_1-penalty promotes a sparse graphical structure in the Ψ_k and by extension $\widehat{\Omega}$. See for example d'Aspremont et al. (2008); Banerjee et al. (2008); Yuan and Lin (2007); Zhou et al. (2011); Zhou (2014) and references therein.

2.1 The projection perspective

The precision matrix (4) has an immediate connection to the K positive-semidefinite Gram matrices $S^k \succeq 0 \in R^{d_k \times d_k}$ associated with each mode of the data tensor $\mathfrak{X}$. For simplicity, we state Lemma 1 for the trace term $\langle \widehat{S}, \Omega_0 \rangle$ in case $n = 1$, with obvious extensions for $n > 1$ and for any $\Omega \in \mathcal{K}_{\mathbf{p}}^{\sharp}$.

Lemma 1 (KS trace: Projection of sample covariances) *Consider the mean zero K-order random tensor $\mathfrak{X} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$. Denote by $X_j^{(k)} \in \mathbb{R}^{d_k}$ the j^{th} column vector in the matrix $\mathbf{X}^{(k)} \in \mathbb{R}^{d_k \times m_k}$ formed by tensor unfolding. Now let $\mathbf{Y}^{(k)} = \mathbf{X}^{(k)T}$. Denote by $Y_i^{(k)} \in \mathbb{R}^{m_k}$ the i^{th} row vector in $\mathbf{X}^{(k)}$. Then for sample covariance $\widehat{S} := \text{vec}\{\, \mathfrak{X}^{\mathrm{T}} \,\} \otimes \text{vec}\{\, \mathfrak{X}^{\mathrm{T}} \,\}$,*

$$\langle \widehat{S}, \Omega_0 \rangle \;\; = \;\; \sum_{k=1}^{K} m_k \, \langle S^k, \Psi_k \rangle \; = \sum_{k=1}^{K} \sum_{j=1}^{m_k} \langle \Psi_k, X_j^{(k)} \otimes X_j^{(k)} \rangle \,, \tag{11}$$

where S^k is the same as in (2)*, and Ω_0 is as in* (5)*. Then we have*

$$\langle \widehat{S}, \Omega_0 \rangle \;\; = \;\; \sum_{k=1}^{K} \sum_{j \neq i}^{d_k} \Psi_{k,ij} \, \langle Y_i^{(k)}, Y_j^{(k)} \rangle \; + \sum_{k=1}^{K} \sum_{i=1}^{d_k} \Psi_{k,ii} \, \langle Y_i^{(k)}, Y_i^{(k)} \rangle \,. \tag{12}$$

Here $\text{vec}\{\, \mathrm{A} \,\}$ *of a matrix $A^{d_k \times m_k}$ is obtained by stacking columns of A into a long vector of size $p = d_k \times m_k$.*

Clearly, (11) holds since

$$\sum_{k=1}^{K} m_k \, \langle S^k, \Psi_k \rangle \;\; = \;\; \sum_{k=1}^{K} \langle \mathbf{X}^{(k)} (\mathbf{X}^{(k)})^T, \Psi_k \rangle \; = \sum_{k=1}^{K} \sum_{j=1}^{m_k} (X_j^{(k)})^T \Psi_k X_j^{(k)};$$

Lemma 1 explains the smoothing ideas. To see this, notice that following (2),

$$\forall i,j \in [d_k], \quad m_k S^k_{ij} \;=\; \langle\, Y^{(k)}_i, Y^{(k)}_j \,\rangle\,, \quad \text{where } Y^{(k)}_j \in \mathbb{R}^{m_k}, \forall j \in [d_k],$$

which in turn can be interpreted as the tensor inner product (1) with $N = K-1$. In particular, when $K=3$, $\langle\, Y^{(k)}_i, Y^{(k)}_j \,\rangle$ corresponds to the inner product of two slices (vectorized) indexed with $i,j \in \{1,\ldots,d_k\}$ respectively. Similarly, we have (12)

$$\sum_{k=1}^K m_k \,\langle\, S^k, \Psi_k \,\rangle \;=\; \sum_{k=1}^K \mathrm{tr}(\mathbf{Y}^{(k)} \Psi_k \mathbf{Y}^{(k)T}) = \sum_{k=1}^K \sum_{i,j=1}^{d_k} \Psi_{k,ij} \,\langle\, Y^{(k)}_i, Y^{(k)}_j \,\rangle\,.$$

Intuitively, we use the m_k fibers to estimate relations between and among the d_k features along the k^{th} mode, as encoded in Ψ_k. Hence, this forms the aggregation of all data from modes other than k, which allows uniform concentration of measure bounds as shown in Lemma 2 to be achieved. In both Greenewald et al. (2019a) and the present work, we focus on error bounds on the estimate of Ω_0 itself, rather than the factors individually.

Lemma 2 *Let $\Delta_\Omega \in \mathcal{K}_{\mathbf{p}}$. Under the conditions in Lemma 1, we have for $\Sigma_0 = \Omega_0^{-1}$,*

$$\frac{\Big|\,\langle\, \mathrm{diag}(\Delta_\Omega), \widehat{S} - \Sigma_0 \,\rangle\,\Big|}{\sqrt{K}\,\|\mathrm{diag}(\Delta_\Omega)\|_F} \;\leq\; C_{\mathrm{diag}}(\max_k \sqrt{d_k})\,\|\Sigma_0\|_2 \left(1 + \max_{k=1}^K \sqrt{\frac{d_k}{m_k}}\right), \tag{13}$$

with probability at least $1 - \sum_{k=1}^K \exp(-cd_k)$, for some absolute constants c, C_{diag}.

We emphasize that we retain essentially the same error bound as that in Greenewald et al. (2019a) for the off-diagonal component of the trace terms in (9). The proof of Lemma 2 appears in Section 5.1. These points will become more clear when we compare the rates of convergence and sample size lower bounds in Theorems 4 and 5. For now, we define the support set of Ω_0.

Definition 3 (The support set of Ω_0) *For each Ψ_k, $k=1,\ldots,K$, denote its support set by* $\mathrm{supp}(\mathrm{offd}(\Psi_k)) = \{(i,j) : i \neq j, \Psi_{k,ij} \neq 0\}$. *Let $s_k := |\mathrm{supp}(\mathrm{offd}(\Psi_k))|$, for all k. Denote the support set of Ω_0 by $\mathcal{S} = \{(i,j) : i \neq j, \Omega_{0,ij} \neq 0\}$, with $s := |\mathcal{S}| = \sum_{k=1}^K m_k s_k$.*

3. The main theorem

We require the following assumptions, where C is an absolute constant.

(A1) Let $\min_k m_k \geq \log p$. Denote by $\delta_{n,k} \asymp \|\Sigma_0\|_2 \sqrt{\frac{\log p}{n m_k}}$ for $k = 1,\ldots,K$. Suppose $\rho_{n,k} = \delta_{n,k}/\varepsilon_k$, where $0 < \varepsilon_k < 1$ for all k.

(A2) The smallest eigenvalue $\phi_{\min}(\Omega_0) = \sum_{k=1}^K \phi_{\min}(\Psi_k) \geq \underline{k}_\Omega > 0$, and the largest eigenvalue $\phi_{\max}(\Omega_0) = \sum_{k=1}^K \phi_{\max}(\Psi_k) \leq \overline{k}_\Omega < \infty$.

(A3) The sample size n and $\min_k m_k =: m_{\min}$ satisfy the following condition:

$$n(m_{\min})^2 \geq C^2(K+1)\kappa(\Sigma_0)^4(s \log p + Kp), \tag{14}$$

where $s = \sum_k m_k s_k$ is the number of non-zero entries in $\mathrm{offd}(\Omega_0)$ as in Definition 3.

Theorem 4 (Main result) *Suppose (A1), (A2), and (A3) hold. Then for absolute constants C, c, c', we have with probability at least $1 - \sum_{k=1}^{K} \exp(-cd_k) - K\exp(-c' \log p)$,*

$$\begin{aligned} \left\|\widehat{\Omega} - \Omega_0\right\|_F / \left\|\Omega_0\right\|_2 &\leq C\kappa(\Sigma_0)\sqrt{\frac{s\log p + Kp}{nm_{\min}}}, \\ \left\|\widehat{\Omega} - \Omega_0\right\|_2 / \left\|\Omega_0\right\|_2 &\leq C\kappa(\Sigma_0)\sqrt{K+1}\sqrt{\frac{s\log p + Kp}{nm_{\min}^2}}, \qquad (15) \\ \textit{and } \left\|\widehat{\Omega} - \Omega_0\right\|_F / \left\|\Omega_0\right\|_F &\leq C\kappa(\Sigma_0)\sqrt{K+1}\sqrt{\frac{s\log p + Kp}{nm_{\min}^2}}. \qquad (16) \end{aligned}$$

The condition number for $\Sigma_0 = \Omega_0^{-1}$ is defined as

$$\kappa(\Sigma_0) = \kappa(\Omega_0) = \left\|\Omega_0\right\|_2 \left\|\Omega_0^{-1}\right\|_2 = \frac{\sum_{k=1}^{K} \phi_{\max}(\Psi_k)}{\sum_{k=1}^{K} \phi_{\min}(\Psi_k)},$$

where we have used the additivity of the eigenvalues of the Kronecker sum. For the sake of comparison, we restate Theorems 1 and 2 therein using order notation in Theorem 5.

Theorem 5 (Greenewald et al. (2019a), restated) *Suppose (A1) and (A2) hold and $n(m_{\min})^2 \geq C^2\kappa(\Sigma_0)^4(s+p)(K+1)^2\log p$, where $s = \sum_k m_k s_k$ is as in Definition 3. Then*

$$\begin{aligned} \frac{\|\widehat{\Omega} - \Omega_0\|_F}{\|\Omega_0\|_2} &= O_p\left(\kappa(\Sigma_0)\sqrt{\frac{(K+1)\,(s+p)\log p}{nm_{\min}}}\right), \qquad (17) \\ \frac{\|\widehat{\Omega} - \Omega_0\|_2}{\|\Omega_0\|_2} &= O_p\left(\kappa(\Sigma_0)(K+1)\sqrt{\frac{(s+p)\log p}{nm_{\min}^2}}\right). \qquad (18) \end{aligned}$$

Summary. The worst aspect ratio is defined as $\max_k(d_k/m_k) = d_{\max}/m_{\min} = p/m_{\min}^2$. Clearly, a smaller aspect ratio implies a faster rate of convergence for the relative errors in the operator and Frobenius norm. Previously in Greenewald et al. (2019a), we provided theoretical guarantees for (9), when the sample size is low, including single-sample convergence when $K \geq 3$. In the present work, these tighter rates in Theorem 4 also expand significantly the range of settings for which TeraLasso is guaranteed to achieve low errors with a constant number of replicates, namely $n = O(1)$, to the $K = 2$ regime. This expansion is important as it closes the gap between the theory in Greenewald et al. (2019a) and the observed $K = 2$ convergence therein, which we now elaborate. First, observe that for (relative) errors in the operator and Frobenius norm in Theorem 5,

$$(K+1)(s+p)\log p \quad \text{therein is replaced with} \quad s\log p + Kp, \quad \text{cf. (16)} \quad \text{and (15)}.$$

Here we eliminate the extraneous $\log p$ factor from the diagonal component of the error through new concentration of measure analysis in the present work; cf. Lemma 2 and the supplementary Lemma 25. This is a significant change for two reasons: (a) note that since p is the product of the d_ks, $\log p = O(\sum_k \log d_k)$ is often nontrivial, especially for larger K; and (b) more importantly, for $K = 2$ and $n = O(1)$ (in contrast to $K > 2$), the error

bound in the operator norm in Theorem 5 from Greenewald et al. (2019a) will *diverge* for any $s \geq 0$ as $p = d_1 d_2$ increases, since

$$\frac{p \log p}{m_{\min}^2} = \frac{d_1 d_2 \log p}{(d_1 \wedge d_2)^2} \geq \log p, \quad \text{where} \quad m_{\min} = p/(d_1 \vee d_2) = d_1 \wedge d_2 \tag{19}$$

and equality holds only when $d_1 = d_2$. As a result, in Theorem 5 Greenewald et al. (2019a), the sample lower bound, namely, $n(m_{\min})^2 \geq C^2 \kappa(\Sigma_0)^4 (s+p)(K+1)^2 \log p$ implies that $n = \Omega(\log p)$, since $m_{\min}^2 \leq p$ in view of (19). In contrast, the lower bound on $n m_{\min}^2$ in (A3) is less stringent, and roughly speaking, saving a factor of $\log p$ so long as $p = \Omega(s \log p)$.

In the present work, under suitable assumptions on the sparsity parameter s and dimension $d_k, \forall k \in [K]$, consistency and the rate of convergence in the operator norm can be obtained so long as $n = \Omega(1)$ and $K \geq 2$. For finite sample settings, namely, when $n = O(1)$, the relative errors will still be bounded at $O_p(1)$ for $K = 2$, for example, when the two dimensions are at the same order: $d_1 \asymp d_2$, and rapidly converge to zero for $K > 2$; cf. Theorem 8. This is consistent with the successful finite sample experiments in both Greenewald et al. (2019a) and the present work, where for $K = 2$, bounded errors in the operator norm are observed as p increases. As a result, our new bound supports the use of the TeraLasso estimator when $K = 2$, so long as a finite number of replicates are available, in a way that the previous Theorem 5 cannot.

Single sample convergence. First of all, both Theorems 4 and 5 imply single sample convergence for the relative error in the operator norm, when $K \geq 3$ and $d_1 \asymp \ldots \asymp d_K$, which we refer to as the cubic tensor settings, since potentially $m_{\min}^2 \geq m_{\min} d_{\max} \log p = p \log p$ will hold. See Theorem 8 in the sequel. However, when the d_ks are skewed, this may not be the case. For example, in our $K = 3$ experiments in the supplementary Section 6 with $d_2 = d_3 = \sqrt{d_1}$, yielding $m_{\min}^2 = p$, our improved result in (15) can provide finite sample convergence in regimes where the bound of (18) from Greenewald et al. (2019a) will not. To make this clear, we first state Corollary 6.

Corollary 6 (Dependence on aspect ratio for $n = 1$) *Suppose (A1, A2) and (A3) hold for $n = 1$. Then with probability at least $1 - K \exp(c \log p)$, we have for some absolute constants c, C,*

$$\left\| \widehat{\Omega} - \Omega_0 \right\|_2 / \|\Omega_0\|_2, \quad \left\| \widehat{\Omega} - \Omega_0 \right\|_F / \|\Omega_0\|_F \leq C K \kappa(\Sigma_0) \sqrt{\frac{d_{\max}}{m_{\min}}} \sqrt{(\log p / K) \sum_{k=1}^{K} \frac{s_k}{d_k} + 1}.$$

Proof Denote by $s = \sum_k m_k s_k$. Then for $n = 1$ and $p = d_{\max} m_{\min} = m_k d_k$ for all k,

$$\begin{aligned} \sqrt{K} \sqrt{\frac{s \log p + Kp}{m_{\min}^2}} &= \sqrt{K \frac{d_{\max}}{m_{\min}}} \sqrt{\frac{\sum_k m_k s_k \log p + Kp}{d_{\max} m_{\min}}} \\ &= \sqrt{K \frac{d_{\max}}{m_{\min}}} \sqrt{\frac{\sum_k m_k s_k \log p + Kp}{p}} = K \sqrt{\frac{d_{\max}}{m_{\min}}} \sqrt{\frac{1}{K} \sum_k \frac{s_k \log p}{d_k} + 1} < 1 \end{aligned}$$

by (A3); The corollary thus follows from (15) and (16) in Theorem 4. ■

Under the bounded aspect ratio regime, the relative errors in the operator and Frobenius norm for estimating the precision matrix Ω_0 depend on the decay of the worst aspect ratio $d_{\max}/m_{\min}$ and the average of $s_k \log p/d_k$ over all modes, which represents relative sparsity levels (sparsity / dimension) in an average sense. For $K > 2$, typically the aspect ratio is much less than 1 and convergence happens rapidly. If the sparse support set is small relative to nominal dimension d_k along each mode, for example, when $\frac{s_k \log p}{d_k} = O(1)$, this convergence is at the rate of decay of the worst aspect ratio. In this case, the diagonal component dominates the rate of convergence and this is essentially optimal, since in the largest component with dimension $d_{\max}$, it has $d_{\max}$ number of parameters to be estimated and $m_{\min} = p/d_{\max}$ number of (hidden) samples for the task. Moreover, K is needed in the bound since we estimate K components all together using one sample in case $n = 1$. We mention in passing that $m_{\min}$ (resp. $nm_{\min}$) appears in the rates of convergence in Theorems 4 and 5 and Lemma 7 as the effective sample size for estimating Ω_0 as a whole for $n = 1$ (resp. $n > 1$). Before we continue, we state Lemma 8 of Greenewald et al. (2019b) in Lemma 7.

Lemma 7 (Lemma 8 of Greenewald et al. (2019b)) *For all* $\Omega \in \mathcal{K}_{\mathbf{p}}$, $\|\Omega\|_2 \leq \sqrt{\frac{K+1}{\min_k m_k}} \|\Omega\|_F$.

Discussions. The proof of Theorem 4 appears in Section 4. Similar properties hold for the relative Frobenius norm error in view of Lemma 7. Note that by (17) and Lemma 7, under the settings of Theorem 5,

$$\left\|\widehat{\Omega} - \Omega_0\right\|_F / \|\Omega_0\|_F = \left\|\widehat{\Omega} - \Omega_0\right\|_F / \|\Omega_0\|_2 \frac{\|\Omega_0\|_2}{\|\Omega_0\|_F} \quad = \quad O_p\left(\kappa(\Sigma_0)(K+1)\sqrt{\frac{(s+p)\log p}{nm_{\min}^2}}\right).$$

As a first example, we consider the **cubic** setting where $d_1 \asymp \ldots \asymp d_K \asymp p^{1/K}$. In words, a tensor is cubical if every mode is about the same size. Then

$$\text{aspect ratio} \quad := \frac{d_{\max}}{m_{\min}} \asymp \frac{p^{1/K}}{p^{1-1/K}} = p^{2/K-1}. \tag{20}$$

Note that for $K > 2$, it becomes possible to obtain a fast rate of convergence in the operator norm for $n = 1$, since in the cubic tensor settings, the effective sample size $m_{\min}$ increases significantly faster than $\sqrt{p}$ given that $d_{\max} = o(p^{1/2})$. We now state Theorem 8, where we consider the cubic tensor setting: d_js are at the same order. More examples are illustrated in the supplementary Section 6.

Theorem 8 (The cubic tensor: $n = 1$**)** *Suppose all conditions in Theorem 4 hold. Suppose* $d_k = O(m_k)$ *for all* $k = 1, \ldots, K$. *Moreover, suppose* $m_1 \asymp m_2 \ldots \asymp m_K$. *Then for* $p = d_{\max} m_{\min}$,

$$\left\|\widehat{\Omega} - \Omega_0\right\|_F / \|\Omega_0\|_2 \quad = \quad O_P\left(\kappa(\Sigma_0)\sqrt{\sum_{k=1}^{K} s_k \log p + K \max_k d_k}\right).$$

Then dense factors are allowed: by (15),

$$\left\|\widehat{\Omega} - \Omega_0\right\|_2 / \|\Omega_0\|_2 = O_P\left(K\kappa(\Sigma_0)\sqrt{\frac{\left(\frac{1}{K}\sum_{k=1}^{K} s_k \log p + d_{\max}\right)}{m_{\min}}}\right), \tag{21}$$

where $d_{\max} \geq d_j \forall j$. *Suppose in addition* $d_{\max} \asymp d_1 \asymp \ldots \asymp d_K = \Omega\left((\log p/K)\sum_k s_k\right)$. *Then*

$$\left\|\widehat{\Omega} - \Omega_0\right\|_2 / \|\Omega_0\|_2, \quad \left\|\widehat{\Omega} - \Omega_0\right\|_F / \|\Omega_0\|_F = O_p\left(K\kappa(\Sigma_0)p^{\frac{1}{K}-\frac{1}{2}}\right). \tag{22}$$

Proof Suppose that $m_1 \asymp m_2 \asymp \ldots \asymp m_K$. Denote by $s = \sum_k m_k s_k$. Then

$$\sqrt{\frac{s\log p + Kp}{(\min_k m_k)}} \;=\; \sqrt{\frac{\sum_k m_k s_k \log p + Kp}{m_{\min}}} \approx \sqrt{K}\sqrt{\frac{1}{K}\sum_k s_k \log p + d_{\max}}$$

The theorem thus follows from Theorem 4. The rest follows from (15), (16), and (20). ■

Optimality of the cubic tensor rate. Theorem 8 shows that convergence will occur for the dense cubic case, so long as

$$m_{\min} = p/d_{\max} = \Omega\big(K\log p\sum_{j=1}^{K} s_j \vee K^2 d_{\max}\big),$$

which is a reasonable assumption in case $K > 2$ and holds under (A3). In other words, the relative errors in the operator and Frobenius norm are bounded so long as the effective sample size $m_{\min}$ is at least $K^2 d_{\max} \geq K\sum_k d_k$, which is roughly K times the total number of (unique) diagonal entries in $\{\Psi_k, k = 1, \ldots, K\}$, and also at least $K\log p$ times $\sum_k s_k$, which in turn denotes the size of total supports $\sum_k |\mathcal{S}_k|$ over off-diagonal components of factor matrices $\{\Psi_1, \ldots, \Psi_k\}$. Consider now an even more special case. Suppose that in the cubic tensor setting, we have in addition $d_{\max} = \Omega(\log p\sum_j s_j/K), \forall j$. Then the error in the operator norm is again dominated by the square root of the aspect ratio parameter. In other words, to achieve the near optimal rate of $O_P(p^{\frac{1}{K}-\frac{1}{2}})$, it is sufficient for each axis dimension $d_k, k \in [K]$ to dominate the *average sparsity* across all factors, namely, $\sum_k s_k/K$ by a $\log p$ factor. Notice that when $d_1 \asymp d_1 \asymp \ldots \asymp d_K = \Omega((\log p/K)\sum_k s_k)$, (21) and (22) follow from Corollary 6 and (20). A more general result has been stated in Corollary 6.

4. Proof of Theorem 4

We focus on the case $n = 1$. For $n > 1$, we defer the proof to Section 4.4. In the proof of Theorem 4 that follows, our strategy will be to show that several events controlling the concentration of the sample covariance matrix (in the $n = 1$ case, simply an outer product) hold with high probability, and then show that given these events hold, the statistical error bounds in Theorem 4 hold. The off-diagonal events are as in (10). We defer the definitions of new diagonal events to Section 5. We use the following notation to describe errors in the precision matrix and its factors. For $\Omega \in \mathcal{K}_{\mathbf{p}}$ let $\Delta_\Omega = \Omega - \Omega_0 \in \mathcal{K}_{\mathbf{p}}$. Since both Ω and Ω_0 are Kronecker sums,

$$\Delta_\Omega \;=\; \Delta_{\Psi_1} \oplus \Delta_{\Psi_2} \oplus \ldots \oplus \Delta_{\Psi_K}$$

for some Δ_{Ψ_k} whose off-diagonal elements are uniquely determined. For an index set S and a matrix $W = [w_{ij}]$, write $W_S \equiv (w_{ij}I((i,j) \in S))$, where $I(\cdot)$ is an indicator function.

4.1 Preliminary results

Before we show the proof of Theorem 4, we need to state the following lemmas. Proofs of Lemmas 7 and 9 appear in Greenewald et al. (2019b) (cf. Lemmas 8 and 10 therein). We then present an error bound for the off-diagonal component of the loss function, which appears as Lemma 12 in Greenewald et al. (2019b) and follows from the concentration of measure bounds on elements of $\text{offd}(S^k - \Sigma_0^{(k)})$; cf. (10). Combined with our new concentration bound on the diagonal component of the loss function, cf. Lemma 2, we obtain the improved overall rate of convergence as stated in Theorem 4.

Lemma 9 *Let $\Omega_0 \succ 0$. Let $S = \{(i,j) : \ \Omega_{0ij} \neq 0, \ i \neq j\}$ and $S^c = \{(i,j) : \ \Omega_{0ij} = 0, \ i \neq j\}$. Then for all $\Delta \in \mathcal{K}_{\mathbf{p}}$, we have*

$$|\Omega_0 + \Delta|_{1,\text{off}} - |\Omega_0|_{1,\text{off}} \quad \geq \quad |\Delta_{S^c}|_1 - |\Delta_S|_1 \tag{23}$$

where by disjointness of $\text{supp}\,(\text{offd}(\Psi_k)) := \{(i,j) : i \neq j, \ \Psi_{k,ij} \neq 0\}, k = 1, \ldots, K,$

$$|\Delta_S|_1 = \sum_{k=1}^{K} m_k \, |\Delta_{\Psi_k,S}|_1 \quad \textit{and} \quad |\Delta_{S^c}|_1 = \sum_{k=1}^{K} m_k \, |\Delta_{\Psi_k,S^c}|_1 \, .$$

Lemma 10 follows from Greenewald et al. (2019b); cf. Lemmas 11 and 12 therein.

Lemma 10 *With probability at least $1 - K\exp(-c'\log p)$, for some absolute constant c'*

$$\left| \, \langle \, \text{offd}(\Delta_\Omega), \widehat{S} - \Sigma_0 \, \rangle \, \right| \leq \sum_{k=1}^{K} m_k \, |\Delta_{\Psi_k}|_{1,\text{off}} \, \delta_k, \qquad \textit{where} \quad \delta_k \asymp \sqrt{\frac{\log p}{m_k}} \, \|\Sigma_0\|_2 \, , \forall k. \tag{24}$$

Next we show that as an immediate corollary of (23), we have Lemma 11, which is a deterministic result and identical to Lemma 10 Greenewald et al. (2019b). The proof is omitted. Lemma 12 follows immediately from Lemmas 10 and 11.

Lemma 11 (Deterministic bounds) *Fix $\rho_k \geq 0$. Denote by*

$$\begin{aligned} \Delta_g \quad &:= \quad \sum_{k=1}^{K} m_k \rho_k \left(|\Psi_k + \Delta_{\Psi_k}|_{1,\text{off}} - |\Psi_k|_{1,\text{off}} \right), \qquad (25) \\ \textit{then} \quad \Delta_g \quad &\geq \quad \sum_{k=1}^{K} m_k \rho_k \left(|\Delta_{\Psi_k,S^c}|_1 - |\Delta_{\Psi_k,S}|_1 \right). \end{aligned}$$

Lemma 12 *Suppose that $d_k = O(m_k)$ for all k. Under the settings of Lemmas 10 and 11, we have for the choice of $\rho_k = \delta_k/\varepsilon_k, \forall k$, where $0 < \varepsilon_k < 1$ and $\delta_k \asymp \sqrt{\frac{\log p}{m_k}} \, \|\Sigma_0\|_2$,*

$$\Delta_g + \ \langle \, \text{offd}(\Delta_\Omega), \widehat{S} - \Sigma_0 \, \rangle \quad \geq \quad -2 \max_k \rho_k \, |\Delta_S|_1 \, , \tag{26}$$

Lemma 13 follows from Lemmas 2 and 12. We provide a proof of Lemma 13 in Section 4.3 for completeness. Since $p = \prod_k d_k \geq 2^K$ so long as $d_k \geq 2$, we have $\log p \geq K$ and hence $\exp(c \log p) > K$ for sufficiently large c. We also note that $\sum_{k=1}^{K} \exp(-cd_k) \leq K\exp(-c' d_{\min}) \leq K\exp(-c'\log p)$ so long as $d_{\min} = \Omega(\log p)$.

Lemma 13 *Suppose that $n = 1$. Let $s = \sum_{k=1}^K m_k s_k$. Let Δ_g be as in Lemma 11. Then, under the settings of Lemmas 2 and 12, we have with probability at least $1-\sum_{k=1}^K \exp(-cd_k)-K\exp(-c'\log p)$,*

$$\left|\Delta_g + \langle \Delta_\Omega, \widehat{S} - \Sigma_0 \rangle\right| \;\leq\; C' \|\Sigma_0\|_2 T_3 \;\; \textit{where } T_3 := \frac{\sqrt{s\log p + Kp}\,\|\Delta_\Omega\|_F}{\sqrt{m_{\min}}}.$$

Proposition 14 *Set $C > 36(\max_k \frac{1}{\varepsilon_k} \vee C_{\rm diag})$ for $C_{\rm diag}$ as in Lemma 22. Let*

$$r_{\mathbf{p}} \;=\; C\|\Sigma_0\|_2\sqrt{s\log p + Kp}/\sqrt{m_{\min}} \quad \textit{where} \quad M = \frac{1}{2}\phi^2_{\max}(\Omega_0) = \frac{1}{2\phi^2_{\min}(\Sigma_0)}. \tag{27}$$

Let $\Delta_\Omega \in \mathcal{K}_{\mathbf{p}}$ such that $\|\Delta_\Omega\|_F = Mr_{\mathbf{p}}$. Then $\|\Delta_\Omega\|_2 \leq \frac{1}{2}\phi_{\min}(\Omega_0)$.

4.2 Proof of Theorem 4

We will only show the proof for $n = 1$. Let

$$G(\Delta_\Omega) \;=\; Q(\Omega_0 + \Delta_\Omega) - Q(\Omega_0) \tag{28}$$

be the difference between the objective function (6) at $\Omega_0 + \Delta_\Omega$ and at Ω_0. Clearly $\widehat{\Delta}_\Omega = \widehat{\Omega} - \Omega_0$ minimizes $G(\Delta_\Omega)$, which is a convex function with a unique minimizer on $\mathcal{K}^\sharp_{\mathbf{p}}$ (cf. Theorem 5 Greenewald et al. (2019b)). Let $r_{\mathbf{p}}$ be as in (27) for some absolute constant C to be specified, and

$$\mathcal{T}_n = \left\{\Delta_\Omega \in \mathcal{K}_{\mathbf{p}} : \Delta_\Omega = \Omega - \Omega_0, \Omega, \Omega_0 \in \mathcal{K}^\sharp_{\mathbf{p}}, \|\Delta_\Omega\|_F = Mr_{\mathbf{p}}\right\}. \tag{29}$$

In particular, we set $C > 36(\max_k \frac{1}{\varepsilon_k} \vee C_{\rm diag})$ in $r_{\mathbf{p}}$, for absolute constant $C_{\rm diag}$ as in Lemma 22. Proposition 15 follows from arguments in Zhou et al. (2010).

Proposition 15 *If $G(\Delta) > 0$ for all $\Delta \in \mathcal{T}_n$ as in (29), then $G(\Delta) > 0$ for all Δ in*

$$\mathcal{V}_n = \{\Delta \in \mathcal{K}_{\mathbf{p}} : \Delta = \Omega - \Omega_0, \Omega, \Omega_0 \in \mathcal{K}^\sharp_{\mathbf{p}}, \|\Delta\|_F > Mr_{\mathbf{p}}\}$$

for $r_{\mathbf{p}}$ (27). Hence if $G(\Delta) > 0$ for all $\Delta \in \mathcal{T}_n$, then $G(\Delta) > 0$ for all $\Delta \in \mathcal{T}_n \cup \mathcal{V}_n$.

Proposition 16 *Suppose $G(\Delta_\Omega) > 0$ for all $\Delta_\Omega \in \mathcal{T}_n$. We then have*

$$\left\|\widehat{\Delta}_\Omega\right\|_F < Mr_{\mathbf{p}}.$$

Proof By definition, $G(0) = 0$, so $G(\widehat{\Delta}_\Omega) \leq G(0) = 0$. Thus if $G(\Delta_\Omega) > 0$ on $\mathcal{T}_n$, then by Proposition 15, $\widehat{\Delta}_\Omega \notin \mathcal{T}_n \cup \mathcal{V}_n$ where $\mathcal{V}_n$ is defined therein. The proposition thus holds. ■

Lemma 17 *Under (A1) - (A3), for all $\Delta \in \mathcal{T}_n$ for which $r_{\mathbf{p}} = o\left(\sqrt{\frac{\min_k m_k}{K+1}}\right)$,*

$$\log|\Omega_0 + \Delta| - \log|\Omega_0| \leq \langle \Sigma_0, \Delta\rangle - \frac{2}{9\|\Omega_0\|_2^2}\|\Delta\|_F^2.$$

Lemma 17 is proved in the supplementary material Section A.2. By Proposition 16, it remains to show that $G(\Delta_\Omega) > 0$ on $\mathcal{T}_n$ under the settings of Lemma 13.

Lemma 18 *With probability at least* $1-\sum_{k=1}^{K}\exp(-cd_k)-K\exp(-c'\log p)$*, we have* $G(\Delta) > 0$ *for all* $\Delta \in \mathcal{T}_n$.

Proof By Lemma 17, if $r_{\mathbf{p}} \leq \sqrt{\min_k m_k/(K+1)}$, we can express (28) as

$$\begin{aligned} G(\Delta_\Omega) &= \langle \Omega_0 + \Delta_\Omega, \widehat{S}\rangle - \log|\Omega_0+\Delta_\Omega| - \langle \Omega_0, \widehat{S}\rangle + \log|\Omega_0| \\ &\quad + \underbrace{\sum_k \rho_k m_k(|\Psi_{k,0}+\Delta_{\Psi,k}|_{1,\text{off}} - |\Psi_{k,0}|_{1,\text{off}})}_{\Delta_g} \qquad (30) \\ &\geq \quad \langle\, \Delta_\Omega, \widehat{S}-\Sigma_0\,\rangle + \frac{2}{9\|\Omega_0\|_2^2}\|\Delta_\Omega\|_F^2 + \Delta_g. \qquad (31) \end{aligned}$$

We have with probability at least $1-\sum_{k=1}^{K}\exp(-cd_k) - K\exp(-c'\log p)$, by (31) and Lemma 13, and for all $\Delta_\Omega \in \mathcal{T}_n$,

$$\begin{aligned} G(\Delta_\Omega) &\geq \frac{2}{9\|\Omega_0\|_2^2}\|\Delta_\Omega\|_F^2 - \left|\Delta_g + \langle\, \Delta_\Omega, \widehat{S}-\Sigma_0\,\rangle\right| \\ &\geq \frac{2}{9\,\|\Omega_0\|_2^2}\,\|\Delta_\Omega\|_F^2 - \frac{C'\,\|\Sigma_0\|_2}{\sqrt{\min_k m_k}}\sqrt{s\log p + Kp}\,\|\Delta_\Omega\|_F =: W, \end{aligned}$$

where $d_k = O(m_k)$ and for $C' = \max_k(\frac{2}{\varepsilon_k}) \vee 2C_{\text{diag}}$,

$$\left|\Delta_g + \langle\, \Delta_\Omega, \widehat{S}-\Sigma_0\,\rangle\right| \quad \leq \quad C'\,\|\Sigma_0\|_2\,\frac{\sqrt{s\log p + Kp}\,\|\Delta_\Omega\|_F}{\sqrt{m_{\min}}}.$$

Now $W > 0$ for $\|\Delta_\Omega\|_F = Mr_{\mathbf{p}}$, where $M = \frac{1}{2\phi^2_{\min}(\Sigma_0)}$, since

$$C'\,\|\Sigma_0\|_2\sqrt{\frac{1}{\min_k m_k}}\sqrt{(Kp+s\log p)}\frac{1}{Mr_{\mathbf{p}}} = \frac{C'}{CM} = \frac{2C'}{C}\phi^2_{\min}(\Sigma_0) < \frac{2}{9\,\|\Omega_0\|_2^2},$$

where we set $C = 18C' = 36(\max_k(\frac{1}{\varepsilon_k}) \vee C_{\text{diag}})$ in $r_{\mathbf{p}}$ as in (27). ■

Theorem 4 follows from Proposition 16; See proof of Lemma 13 for the error probability. The error in the operator norm follows from that of the Frobenius norm and Lemma 7. □

4.3 Proof of Lemma 13 and Proposition 14

Proof of Lemma 13. We focus on the case $d_k \leq m_k \forall k$; By definition of Δ_g,

$$\langle\, \Delta, S-\Sigma_0\,\rangle + \Delta_g := \langle\, \text{offd}(\Delta), S-\Sigma_0\,\rangle + \Delta_g + \langle\, \text{diag}(\Delta), S-\Sigma_0\,\rangle$$

Combining Lemmas 2 and 10, we have with probability at least

$$1-\sum_{k=1}^{K}\exp(-cd_k) - K\exp(-c'\log p),$$

for $d_k = O(m_k)$ and $|\Delta_S|_1 \le \sqrt{s}\,\|\Delta_S\|_F$,

$$\begin{aligned}
\Big|\Delta_g + \langle \operatorname{offd}(\Delta_\Omega), \widehat{S} - \Sigma_0 \rangle \Big| &\le 2 \max_k \rho_k \,|\Delta_S|_1 \le 2 \max_k \big(\frac{1}{\varepsilon_k}\sqrt{\frac{\log p}{m_k}}\big)\sqrt{s}\,\|\Delta_{\Omega,S}\|_F \\
\text{and } \Big|\langle \operatorname{diag}(\Delta_\Omega), \widehat{S} - \Sigma_0 \rangle \Big| &\le C_{\text{diag}}\,\|\Sigma_0\|_2\,\sqrt{d_{\max}}\sqrt{K}\,\|\operatorname{diag}(\Delta_\Omega)\|_F \left(1 + \sqrt{\frac{d_{\max}}{m_{\min}}}\right),
\end{aligned}$$

where $s = \sum_{k=1}^K m_k s_k$; cf. (26) and (13). The Lemma thus holds by the triangle inequality: for $d_{\max} \le \sqrt{p}$

$$\begin{aligned}
|\langle \Delta, S - \Sigma_0 \rangle + \Delta_g| &\le \Big|\langle \operatorname{offd}(\Delta), \widehat{S} - \Sigma_0 \rangle + \Delta_g\Big| + \Big|\langle \operatorname{diag}(\Delta), \widehat{S} - \Sigma_0 \rangle\Big| \\
&\le 2C_{\text{offd}}\,\|\Sigma_0\|_2 \sqrt{\frac{s\log p}{m_{\min}}}\,\|\Delta_{\Omega,S}\|_F + C_{\text{diag}}\,\|\Sigma_0\|_2\,\sqrt{d_{\max}}\sqrt{K}\,\|\operatorname{diag}(\Delta_\Omega)\|_F \left(1 + \sqrt{\frac{d_{\max}}{m_{\min}}}\right) \\
&\le 2C_{\text{offd}} \vee C_{\text{diag}}\,\|\Sigma_0\|_2 \big(\sqrt{\frac{s\log p}{m_{\min}}}\,\|\Delta_{\Omega,S}\|_F + \sqrt{K}\,\|\operatorname{diag}(\Delta_\Omega)\|_F\,\frac{\sqrt{p} + d_{\max}}{2\sqrt{m_{\min}}}\big) \\
&\le C'\,\|\Sigma_0\|_2\,T_3,
\end{aligned}$$

where $C_{\text{offd}} \asymp \max_k \big(1/\varepsilon_k\big)$, $C' = 2(C_{\text{diag}} \vee C_{\text{offd}})$, and by Cauchy-Schwarz inequality,

$$\big(\sqrt{s\log p}\,\|\operatorname{offd}(\Delta_\Omega)\|_F + \sqrt{Kp}\,\|\operatorname{diag}(\Delta_\Omega)\|_F\big) \le \sqrt{s\log p + pK}\,\|\Delta_\Omega\|_F\,.$$

□

Proof of Proposition 14. Indeed, by Theorem 7, we have for all $\Delta \in \mathcal{T}_n$,

$$\begin{aligned}
\|\Delta\|_2 &\le \sqrt{\frac{K+1}{\min_k m_k}}\|\Delta\|_F = \sqrt{\frac{K+1}{m_{\min}}}\,M r_{\mathbf{p}} \\
&\le \sqrt{\frac{K+1}{m_{\min}}}\frac{C}{2}\frac{1}{\phi^2_{\min}(\Sigma_0)}\,\|\Sigma_0\|_2\sqrt{\frac{s\log p + pK}{m_{\min}}} \le \frac{1}{2}\phi_{\min}(\Omega_0) = \frac{1}{2\phi_{\max}(\Sigma_0)},
\end{aligned}$$

so long as $m^2_{\min} > 2C^2(K+1)\kappa(\Sigma_0)^4(s\log p + pK)$. □

4.4 Extension to multiple samples $n > 1$

To complete the proof, it remains to present the case of $n > 1$. Incorporating $n > 1$ directly into the proof above is relatively straightforward but notation-dense; hence it suffices to note that having n independent samples essentially increases the m_k replication to nm_k, and propagate this fact through the proof. We also note that the multi-sample $n > 1$ case can be converted to the single sample $n = 1$ regime to obtain a result directly. To see this, note that n independent samples with precision matrix $\Omega_0 \in \mathbb{R}^{p\times p}$ can be represented as a single sample with the block-diagonal precision matrix, i.e. Ω_0 repeated n times blockwise along the diagonal, specifically, $\Omega^{(n)} = I_n \otimes \Omega_0 \in \mathbb{R}^{pn\times pn}$. Recall that by definition of the Kronecker sum, $\Omega^{(n)} = I_n \otimes \Omega_0 = 0_{n\times n} \oplus \Psi_1 \oplus \cdots \oplus \Psi_K$ is a $(K+1)$-order Kronecker sum with $p^{(n)} = pn$, achieved by introducing an all-zero factor $\Psi_0 = 0_{n\times n}$ with $d_0 = n$ (and $m_0 = p$). Since this extra factor is zero, the operator norms are not affected. The sparsity

factor of $\Omega^{(n)}$ is $s^{(n)} = sn$ since the non-zero elements are replicated n times, and each co-dimension $m_k^{(n)} := p^{(n)}/d_k = nm_k$ for $k > 0$. Hence the single sample convergence result can be applied with $K^{(n)} = K + 1$, yielding for $n \leq d_{\max}$ and $K \geq 2$

$$\begin{aligned}\left\|\widehat{\Omega} - \Omega_0\right\|_2 / \left\|\Omega_0\right\|_2 &= C\kappa(\Sigma_0)\sqrt{K^{(n)}+1}\sqrt{\frac{s^{(n)}\log p^{(n)} + K^{(n)}p^{(n)}}{[m_{\min}^{(n)}]^2}} \\ &= C\kappa(\Sigma_0)\sqrt{K+2}\sqrt{\frac{s(\log p + \log n) + (K+1)p}{nm_{\min}^2}} \\ &\leq C\sqrt{\frac{8}{3}}\kappa(\Sigma_0)\sqrt{K+1}\sqrt{\frac{s\log p + Kp}{nm_{\min}^2}}\end{aligned}$$

since $m_{\min}^{(n)} = \min(m_0, nm_{\min}) = \min(d_{\max}m_{\min}, nm_{\min}) = nm_{\min}$ whenever $n \leq d_{\max}$. Hence Theorem 4 is recovered for $n \leq d_{\max}$.

5. The new concentration bounds on the diagonal

Let $c, C, C_1, C_{\text{diag}}, \ldots$ be absolute constants which may change line by line. We will restate Lemma 2 in Lemma 22, which is used to replace Lemma 13 Greenewald et al. (2019a) in the proof of the main Theorem 4. In order to do so, we need to set up some notation. Recall, when we have K-way tensor, $X_j^{(k)}$ is a fiber (column vector) in the matrix formed by tensor unfolding of $\mathfrak{X}$ along the k^{th}-mode, denoted by $\mathbf{X}^{(k)}$, formed by rearranging the mode-k fibers as columns of the resulting matrix of dimension $d_k \times m_k$ where $m_k = p/d_k$. Now let $\mathbf{Y}^{(k)} = (\mathbf{X}^{(k)})^T$. First we show the following bounds on the ε-net of $\mathbb{S}^{d_k-1}, \forall k$, as constructed in Lemma 19; See for example Milman and Schechtman (1986).

Lemma 19 *Let $1/2 > \varepsilon > 0$. For each $k \in [K]$, there exists an ε-net Π_{d_k}, which satisfies*

$$\Pi_{d_k} \subset \mathbb{S}^{d_k-1} \quad \textit{and} \quad |\Pi_{d_k}| \leq (1 + 2/\varepsilon)^{d_k}.$$

The proof of Lemma 20 appears in Section 5.3.

Lemma 20 *Let $\mathbb{S}^{d_k-1}$ be the sphere in $\mathbb{R}^{d_k}$, and a vector $\delta = (\delta_1, \ldots, \delta_{d_k}) \in \mathbb{S}^{d_k-1}$. Construct an ε-net $\Pi_{d_k} \subset \mathbb{S}^{d_k-1}$ such that $|\Pi_{d_k}| \leq (1 + 2/\varepsilon)^{d_k}$. Denote by $t_{\delta,k}$ a chosen parameter as follows:*

$$t_{\delta,k} := C_m\sqrt{p}\left\|\Sigma_0\right\|_2\left(1 \vee \sqrt{\frac{d_k}{m_k}}\frac{\left\|\delta\right\|_\infty}{\left\|\delta\right\|_2}\right), \tag{32}$$

where for a vector $\delta \in \mathbb{R}^{d_k}$, we have $\left\|\delta\right\|_\infty = \max_j |\delta_j|$. Define the event $\mathcal{G}_k$ as:

$$\textit{event}\ \ \mathcal{G}_k = \left\{\sup_{\delta\in\Pi_{d_k}}\sum_{i=1}^{d_k}\delta_i\left(\left\langle Y_i^{(k)}, Y_i^{(k)}\right\rangle - \mathbb{E}\left\langle Y_i^{(k)}, Y_i^{(k)}\right\rangle\right) \leq t_{\delta,k}\right\},$$

where recall $\mathbf{Y}^{(k)} = (\mathbf{X}^{(k)})^T$. *Moreover, we have by a standard approximation argument and the union bound, on event* $\mathcal{G} = \mathcal{G}_1 \cap \ldots \cap \mathcal{G}_K$,

$$\forall k \quad \sup_{\delta \in \mathbb{S}^{d_k - 1}} \sum_{i=1}^{d_k} \delta_i \left(\langle Y_i^{(k)}, Y_i^{(k)} \rangle - \mathbb{E} \langle Y_i^{(k)}, Y_i^{(k)} \rangle \right) \leq \frac{t_{\delta,k}}{1-\varepsilon}.$$

Finally, $\mathbb{P}(\mathcal{G}) \geq 1 - \sum_k \exp(-cd_k)$ *for some absolute constant* c.

Moreover, we need the following event $\mathcal{D}_0$:

$$\text{event} \;\; \mathcal{D}_0 = \left\{ \left| \langle I_p, \widehat{S} - \Sigma_0 \rangle \right| \leq C\sqrt{p} \, \|\Sigma_0\|_2 \sqrt{\log p / n} \right\},$$

which states that the sum of the errors of the diagonal of the sample covariance $\widehat{S}$ (7) is bounded tightly in an average sense. Indeed, $\mathrm{tr}(\widehat{S})$ converges to $\mathrm{tr}(\Sigma_0)$ at the rate of

$$\left| \mathrm{tr}(\widehat{S}) - \mathrm{tr}(\Sigma_0) \right| / p = O_P(\|\Sigma_0\|_2 \sqrt{\log p / (np)}).$$

Finally, we define the unified event $\mathcal{A}$ as the event that all these events hold, i.e.

$$\mathcal{A} = \mathcal{T} \cap \mathcal{D}_0 \cap \mathcal{G}, \quad \text{where} \quad \mathcal{G} = \mathcal{G}_1 \cap \cdots \cap \mathcal{G}_K,$$

whose probability of holding will yield the probability as stated in the main Theorem 4.

5.1 Proof of Lemma 2

Throughout this proof, we assume $n = 1$ for simplicity. We now provide outline for proving the upper bound on the diagonal component of the main result of the paper. By Lemma 1, we have for the diagonal and off-diagonal components of the trace term defined as follows:

$$\begin{aligned} \langle \widehat{S}, \mathrm{diag}(\Psi_1) \oplus \cdots \oplus \mathrm{diag}(\Psi_K) \rangle &= \sum_{k=1}^{K} \sum_{i=1}^{d_k} \Psi_{k,ii} \langle Y_i^{(k)}, Y_i^{(k)} \rangle, \qquad (33) \\ \langle \widehat{S}, \mathrm{offd}(\Psi_1) \oplus \cdots \oplus \mathrm{offd}(\Psi_K) \rangle &= \sum_{k=1}^{K} \sum_{i \neq j} \Psi_{k,ij} \langle Y_i^{(k)}, Y_j^{(k)} \rangle. \end{aligned}$$

where recall the true parameter $\Omega_0 = \Psi_1 \oplus \cdots \oplus \Psi_K$, where $\Psi_k \in \mathbb{R}^{d_k \times d_k}$ (4). See also (9). Let $\Omega \in \mathcal{K}_{\mathbf{p}}$. Since both Ω and Ω_0 are Kronecker sums,

$$\Delta_\Omega \;\; := \;\; \Omega - \Omega_0 = \Delta_{\Psi_1} \oplus \Delta_{\Psi_2} \oplus \ldots \oplus \Delta_{\Psi_K}, \qquad (34)$$

for some $\Delta_{\Psi_k} \in \mathbb{R}^{d_k \times d_k}$ whose off-diagonal elements are uniquely determined. The existence of such parameterization is given in Lemma 7 Greenewald et al. (2019b). For self-containment, we state Lemma 21, where we also state the notation we use throughout this section.

Lemma 21 (Decomposition lemma on Δ_Ω) *Let* $\Omega \in \mathcal{K}_{\mathbf{p}}$. *Then* $\Delta_\Omega = \Omega - \Omega_0 \in \mathcal{K}_{\mathbf{p}}$. *To obtain a uniquely determined representation, as in Greenewald et al. (2019b), we can rewrite* (34) *as follows:*

$$\begin{aligned} \Delta_\Omega &= \Delta'_\Omega + \tau_\Omega I_p, \quad \textit{where} \quad \tau_\Omega = \mathrm{tr}(\Delta_\Omega)/p, \quad \textit{and} \\ \Delta'_\Omega &= \Delta'_{\Psi_1} \oplus \ldots \oplus \Delta'_{\Psi_K}, \quad \textit{where} \quad \mathrm{tr}(\Delta'_{\Psi_k}) = 0, \forall k; \qquad (35) \end{aligned}$$

Here we use the trace-zero convention which guarantees the uniqueness of the Δ'_{Ψ_k}.

The off-diagonal component has been dealt with in Lemma 10. Hence in the rest of this section, we present the key idea to the improved rates in Theorem 4, by proving Lemma 22. Denote by

$$\mathrm{diag}(\Delta'_{\Psi_k}) = \mathrm{diag}(\delta_1^k, \ldots, \delta_{d_k}^k) =: \mathrm{diag}(\delta^k) \quad \text{where} \quad m_k = p/d_k, \quad \forall k \tag{36}$$

and Δ'_{Ψ_k} is the same as in (35) and (39). We now restate Lemma 2 in Lemma 22.

Lemma 22 (New diagonal bound) *Following the notation as in Lemma 21, we have on event $\mathcal{D}_0 \cap \mathcal{G}$, which holds with probability at least $1 - \sum_k \exp(-cd_k) - 1/p^4$,*

$$\begin{aligned}
\left| \langle\, \mathrm{diag}(\Delta_\Omega), \widehat{S} - \Sigma_0 \,\rangle \right| &\leq C_1 \left\|\Sigma_0\right\|_2 \left(\sqrt{\sum_{k=1}^K m_k \left\| \mathrm{diag}(\Delta'_{\Psi_k}) \right\|_F^2 + \tau_\Omega^2 p} \sqrt{\sum_{k=1}^K d_k + \log p} \right) \\
&\quad + C_0 \left\|\Sigma_0\right\|_2 \left(\sum_{k=1}^K d_k \left\| \mathrm{diag}(\Delta'_{\Psi_k}) \right\|_2 \right) \\
\textit{and hence} \quad \frac{\left| \langle\, \mathrm{diag}(\Delta_\Omega), \widehat{S} - \Sigma_0 \,\rangle \right|}{\sqrt{K} \left\| \mathrm{diag}(\Delta_\Omega) \right\|_F} &\leq C_2 \max_k \sqrt{d_k} \left\|\Sigma_0\right\|_2 \left(1 + \max_k \sqrt{\frac{d_k}{m_k}} \right),
\end{aligned}$$

where c, C_0, C_1, C_2 are absolute constants.

Note when we have $d_k \leq \sqrt{p}$ for all k, or equivalently, when $\max_k \sqrt{\frac{d_k}{m_k}} \leq 1$, we do not need to pay the extra factor of $\sqrt{\log p}$ as in Lemma 13 Greenewald et al. (2019b) on the diagonal portion of the error bound, resulting in the improved rates of convergence in Theorem 4. When $d_k = o(m_k \log p), \forall k$, the bound in Lemma 22 still leads to an improvement on the overall rate.

5.2 Proof of Lemma 22

Let the sample covariance $\widehat{S} := \mathrm{vec}\{\, \mathcal{X}^{\mathrm{T}} \,\} \otimes \mathrm{vec}\{\, \mathcal{X}^{\mathrm{T}} \,\}$ be as in (7) and $\Sigma_0 = \Omega_0^{-1} \in \mathbb{R}^{n \times n}$ be the true covariance matrix. Thus we denote by $\mathrm{vec}\{\, \mathcal{X}^{\mathrm{T}} \,\} \sim \Sigma_0^{1/2} \mathrm{Z}$, where $Z \in \mathbb{R}^p$ denotes an isotropic subgaussian random vector with independent coordinates. Let

$$\mathrm{diag}(\widetilde{\Delta}_k) \quad := \quad I_{[d_{1:k-1}]} \otimes \mathrm{diag}(\Delta'_{\Psi_k}) \otimes I_{[d_{k+1:K}]}. \tag{37}$$

Consequently, $\left\| \mathrm{diag}(\widetilde{\Delta}_k) \right\|_F^2 := m_k \left\| \mathrm{diag}(\Delta'_{\Psi_k}) \right\|_F^2$. Lemma 23 follows from Lemma 21.

Lemma 23 (Decomposition on the diagonal) *Under the settings of Lemma 21, we have $\Delta'_\Omega = \Delta_\Omega - \tau_\Omega I_p$, with the following expression on its diagonal component:*

$$\begin{aligned}
\mathrm{diag}(\Delta'_\Omega) &= \mathrm{diag}(\Delta'_{\Psi_1}) \oplus \mathrm{diag}(\Delta'_{\Psi_2}) \oplus \ldots \oplus \mathrm{diag}(\Delta'_{\Psi_K}) \\
&= \sum_{k=1}^K I_{[d_{1:k-1}]} \otimes \mathrm{diag}(\Delta'_{\Psi_k}) \otimes I_{[d_{k+1:K}]} =: \sum_{k=1}^K \mathrm{diag}(\widetilde{\Delta}_k),
\end{aligned} \tag{38}$$

where $\tau_\Omega = \mathrm{tr}(\Delta_\Omega)/p$ *and* $\mathrm{tr}(\Delta'_{\Psi_j}) = 0, \forall j$. *Then by orthogonality of the decomposition,*

$$\|\mathrm{diag}(\Delta_\Omega)\|_F^2 = \sum_{k=1}^K m_k \left\|\mathrm{diag}(\Delta'_{\Psi_k})\right\|_F^2 + p\tau_\Omega^2. \tag{39}$$

Moreover, we have for $\sqrt{d_{\max}/m_{\min}} := \left(\max_{k=1}^K \sqrt{d_k/m_k}\right)$,

$$\sum_{k=1}^K \sqrt{d_k}\left\|\mathrm{diag}(\Delta'_{\Psi_k})\right\|_2 \;\leq\; \sqrt{d_{\max}/m_{\min}}\sqrt{K}\,\|\mathrm{diag}(\Delta_\Omega)\|_F\,. \tag{40}$$

Proof of Lemma 23. First, (39) follows from Lemma 7 Greenewald et al. (2019b). Now we have by elementary inequalities:

$$\begin{aligned}\sum_{k=1}^K \sqrt{d_k}\left\|\mathrm{diag}(\Delta'_{\Psi_k})\right\|_2 &= \sum_{k=1}^K \sqrt{\frac{d_k}{m_k}}\sqrt{m_k}\left\|\mathrm{diag}(\Delta'_{\Psi_k})\right\|_2 \\ &\leq \sqrt{d_{\max}/m_{\min}}\sqrt{K}\sqrt{\sum_{k=1}^K m_k\left\|\mathrm{diag}(\Delta'_{\Psi_k})\right\|_2^2}.\end{aligned}$$

Thus (40) holds in view of (39). □

Proof of Lemma 22. For the $t_{\delta,k}$ in (32), we have

$$\begin{aligned}t_{\delta,k}\left\|\delta^k\right\|_2 &= \|\Sigma_0\|_2\sqrt{p}\left\|\delta^k\right\|_2\left(1 \vee \frac{d_k}{\sqrt{p}}\frac{\left\|\delta^k\right\|_\infty}{\|\delta^k\|_2}\right) \\ &= \|\Sigma_0\|_2\left(\sqrt{d_k m_k}\left\|\mathrm{diag}(\Delta'_{\Psi_k})\right\|_F \vee d_k\left\|\mathrm{diag}(\Delta'_{\Psi_k})\right\|_2\right), \\ \text{where}\quad \left\|\delta^k\right\|_2 &:= \left\|\mathrm{diag}(\Delta'_{\Psi_k})\right\|_F \quad\text{and}\quad \left\|\delta^k\right\|_\infty = \left\|\mathrm{diag}(\Delta'_{\Psi_k})\right\|_2.\end{aligned}$$

We use a shorthand notation: for some absolute constant C_m,

$$t'_k \;:=\; C_m\,\|\Sigma_0\|_2\left(\sqrt{m_k d_k}\left\|\mathrm{diag}(\Delta'_{\Psi_k})\right\|_F \vee d_k\left\|\mathrm{diag}(\Delta'_{\Psi_k})\right\|_2\right). \tag{41}$$

First, we have by (36) and (37), using the expression in (33),

$$\begin{aligned}\langle\,\mathrm{diag}(\widetilde{\Delta}_k), \widehat{S} - \Sigma_0\,\rangle &= \sum_{k=1}^K m_k\,\langle\, S^{(k)} - \mathbb{E}S^{(k)}, \mathrm{diag}(\Delta'_{\Psi_k})\,\rangle \\ &= \mathrm{tr}(\mathbf{Y}^{(k)}\mathrm{diag}(\delta_1^k,\ldots,\delta_{d_k}^k)\mathbf{Y}^{(k)T}) - \mathbb{E}\mathrm{tr}(\mathbf{Y}^{(k)}\mathrm{diag}(\delta_1^k,\ldots,\delta_{d_k}^k)\mathbf{Y}^{(k)T}) \\ &= \sum_{i=1}^{d_k}\delta_i^k\left(\,\langle\, Y_i^{(k)}, Y_i^{(k)}\,\rangle\; - \mathbb{E}\,\langle\, Y_i^{(k)}, Y_i^{(k)}\,\rangle\,\right),\end{aligned}$$

where it is understood that $\delta^k \in \mathbb{R}^{d_k}$ is associated with Δ'_{Ψ_k} through:

$$\mathrm{diag}(\Delta'_{\Psi_k}) = \mathrm{diag}(\delta_1^k,\ldots,\delta_{d_k}^k).$$

Now, on event $\mathcal{G}$, using the expressions in (33), we have by Lemma 20, simultaneously for all Δ'_Ω as defined in (39),

$$\begin{aligned}
&\left| \langle \operatorname{diag}(\Delta'_\Omega), \widehat{S} - \Sigma_0 \rangle \right| = \left| \sum_{k=1}^{K} \langle \operatorname{diag}(\widetilde{\Delta}_k), \widehat{S} - \Sigma_0 \rangle \right| \\
&\leq \sum_k \sup_{\delta^k \in \mathbb{R}^{d_k}} \sum_{i=1}^{d_k} \delta_i^k \left(\langle Y_i^{(k)}, Y_i^{(k)} \rangle - \mathbb{E} \langle Y_i^{(k)}, Y_i^{(k)} \rangle \right) \leq \sum_k \frac{t_{\delta,k} \left\| \delta^k \right\|_2}{1-\varepsilon} =: \sum_k t'_k.
\end{aligned}$$

Moreover, under $\mathcal{D}_0$, we have for $n = 1$,

$$\tau_\Omega \left| \langle I_p, \widehat{S} - \Sigma_0 \rangle \right| \leq C_1 \tau_\Omega \sqrt{p} \left\| \Sigma_0 \right\|_2 \sqrt{\log p}.$$

Finally, summing these terms together and using (40), we have on $\mathcal{G} \cap \mathcal{D}_0$,

$$\begin{aligned}
&\left| \langle \operatorname{diag}(\Delta_\Omega), \widehat{S} - \Sigma_0 \rangle \right| \leq \left| \langle \tau_p I_p, \widehat{S} - \Sigma_0 \rangle \right| + \left| \langle \operatorname{diag}(\Delta'_\Omega), \widehat{S} - \Sigma_0 \rangle \right| \\
&\leq C_0 \left\| \Sigma_0 \right\|_2 \left(\tau_\Omega \sqrt{p} \sqrt{\log p} + \sum_{k=1}^{K} \sqrt{d_k} \sqrt{m_k} \left\| \operatorname{diag}(\Delta'_{\Psi_k}) \right\|_F + \sum_{k=1}^{K} d_k \left\| \operatorname{diag}(\Delta'_{\Psi_k}) \right\|_2 \right) =: r_{\text{diag}}.
\end{aligned}$$

It remains to bound r_{diag}. First, we have by the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\tau_\Omega \sqrt{p} \sqrt{\log p} + \sum_{k=1}^{K} \sqrt{d_k} \sqrt{m_k} \left\| \operatorname{diag}(\Delta'_{\Psi_k}) \right\|_F \\
&\leq \sqrt{\log p + \sum_{k=1}^{K} d_k} \sqrt{\sum_{k=1}^{K} m_k \left\| \operatorname{diag}(\Delta'_{\Psi_k}) \right\|_F^2 + \tau_\Omega^2 p} \leq c \sqrt{\sum_{k=1}^{K} d_k} \left\| \operatorname{diag}(\Delta_\Omega) \right\|_F,
\end{aligned}$$

and clearly $\log p = \sum_{k=1} \log d_k \leq \sum_{k=1}^{K} d_k$, since the RHS is a polynomial function of p, and the last line holds by (39). Moreover, we have by (40),

$$\sum_{k=1}^{K} d_k \left\| \operatorname{diag}(\Delta'_{\Psi_k}) \right\|_2 \leq \max_k \frac{d_k}{\sqrt{m_k}} \sqrt{K} \left\| \operatorname{diag}(\Delta_\Omega) \right\|_F.$$

Combining the two bounds immediately above, we have

$$\begin{aligned}
r_{\text{diag}} &\leq C_0 \left\| \Sigma_0 \right\|_2 \left(\left\| \operatorname{diag}(\Delta_\Omega) \right\|_F \sqrt{\sum_{k=1}^{K} d_k} + \max_k \frac{d_k}{\sqrt{m_k}} \sqrt{K} \left\| \operatorname{diag}(\Delta_\Omega) \right\|_F \right) \\
&\leq \left\| \Sigma_0 \right\|_2 (\max_k \sqrt{d_k}) \sqrt{K} \left\| \operatorname{diag}(\Delta_\Omega) \right\|_F \left(1 \vee \max_{k=1}^{K} \sqrt{\frac{d_k}{m_k}} \right).
\end{aligned}$$

Set $t = C\sqrt{p \log p} \left\| \Sigma_0 \right\|_2$. Next we show that $\mathbb{P}(\mathcal{D}_0) \geq 1 - \frac{1}{p^4}$. To bound $\mathcal{D}_0$, we rewrite

$$\langle \widehat{S} - \Sigma_0, I \rangle = \operatorname{tr}(\widehat{S} - \Sigma_0) = Z^T \Sigma_0 Z - \mathbb{E}(Z^T \Sigma_0 Z),$$

where $\Sigma_0 \succ 0$ is a $p \times p$ symmetric positive definite matrix and $Z \in \mathbb{R}^p$ is the same as in (3). Thus, we have by the Hanson-Wright inequality,

$$\begin{aligned}
&\mathbb{P}\left(\left| \langle \widehat{S} - \Sigma_0, I \rangle \ > C \left\| \Sigma_0 \right\|_2 \sqrt{p \log p} \right| \right) \\
&= \ \mathbb{P}\left(\left| Z^T \Sigma_0 Z - E Z^T \Sigma_0 Z \right| > C \sqrt{p \log p} \left\| \Sigma_0 \right\|_2 \right) \\
&\leq \ 2 \exp\left(-c \min\left(\frac{C^2 p \log p \left\| \Sigma_0 \right\|_2^2}{\left\| \Sigma_0 \right\|_F^2}, C \sqrt{p \log p} \right) \right) \leq \frac{1}{p^4},
\end{aligned}$$

where $(C^2 \wedge C)c \geq 4$ and $\left\| \Sigma_0 \right\|_F \leq \sqrt{p} \left\| \Sigma_0 \right\|_2$. Hence by the union bound, we have by Lemma 20 and the bound on $\mathcal{D}_0$ immediately above,

$$\mathbb{P}\left(\mathcal{G} \cap \mathcal{D}_0 \right) \geq 1 - c \exp(-\log p) - \sum_k \exp(-c d_k).$$

The lemma thus holds upon adjusting the constants. □

5.3 Proof of Lemma 20

Let $\delta \in \mathbb{R}^{d_k}$. Using the notation in Lemma 23, let $\mathrm{diag}(\Delta'_{\Psi_k}) = \mathrm{diag}(\delta_1, \ldots, \delta_{d_k})$, and for each k,

$$\mathrm{diag}(\widetilde{\Delta}_k) := I_{[d_{1:k-1}]} \otimes \mathrm{diag}(\Delta'_{\Psi_k}) \otimes I_{[d_{k+1:K}]} \in \mathbb{R}^{p \times p}.$$

Denote by $Y_j^{(k)} \in \mathbb{R}^{m_k}, j = 1, \ldots, d_k$ the j^{th} row vector in $\mathbf{X}^{(k)} \in \mathbb{R}^{d_k \times m_k}$. Recall $\mathbf{Y}^{(k)} = (\mathbf{X}^{(k)})^T$, where $\mathbf{X}^{(k)}$ is a matrix of dimension $d_k \times m_k$. Recall that we have for each $k \in [K]$,

$$\begin{aligned}
&\left| \langle \mathrm{diag}(\widetilde{\Delta}_k), \widehat{S} - \Sigma_0 \rangle \right| = \left| \langle \mathrm{diag}(\widetilde{\Delta}_k), \Sigma_0^{1/2} Z Z^T \Sigma_0^{1/2} - \Sigma_0 \rangle \right| \\
&= \ \langle \mathrm{vec}\{ \mathbf{Y}^{(\mathrm{k})} \} \otimes \mathrm{vec}\{ \mathbf{Y}^{(\mathrm{k})} \} - \mathbb{E}\mathrm{vec}\{ \mathbf{Y}^{(\mathrm{k})} \} \otimes \mathrm{vec}\{ \mathbf{Y}^{(\mathrm{k})} \}, \mathrm{diag}(\Delta'_{\Psi_{\mathrm{k}}}) \otimes \mathrm{I}_{\mathrm{m_k}} \rangle \\
&= \ \sum_{j=1}^{d_k} \delta_j \left(\langle Y_j^{(k)}, Y_j^{(k)} \rangle \ - \mathbb{E} \langle Y_j^{(k)}, Y_j^{(k)} \rangle \right).
\end{aligned}$$

Moreover for a vector $\delta = (\delta_1, \ldots, \delta_{d_k}) \in \mathbb{R}^{d_k}$ and $\mathrm{diag}(\Delta'_{\Psi_k}) = \mathrm{diag}(\delta_1, \ldots, \delta_{d_k})$, it holds that

$$t_{\delta,k} \|\delta\|_2 \ = \ C_m \left\| \Sigma_0 \right\|_2 \left(\sqrt{p} \left\| \mathrm{diag}(\Delta'_{\Psi_k}) \right\|_F \vee d_k \left\| \mathrm{diag}(\Delta'_{\Psi_k}) \right\|_2 \right) =: t'_k. \tag{42}$$

To bound the probability for event $\mathcal{G}_k$, we first rewrite the trace of $\mathrm{diag}(\widetilde{\Delta}_k)$ and $\widehat{S} - \Sigma_0$:

$$\left| \langle \mathrm{diag}(\widetilde{\Delta}_k), \widehat{S} - \Sigma_0 \rangle \right| \ = \ \left| Z^T \Sigma_0^{1/2} \mathrm{diag}(\widetilde{\Delta}_k) \Sigma_0^{1/2} Z - \mathbb{E}(Z^T \Sigma_0^{1/2} \mathrm{diag}(\widetilde{\Delta}_k) \Sigma_0^{1/2} Z) \right|.$$

Then we have by the Hanson-Wright inequality Rudelson and Vershynin (2013), for $t_{\delta,k}$ as defined in (42),

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{i=1}^{d_k}\frac{\delta_i}{\|\delta\|_2}\left(\langle Y_i^{(k)}, Y_i^{(k)}\rangle - \mathbb{E}\langle Y_i^{(k)}, Y_i^{(k)}\rangle\right)\right| \geq t_{\delta,k}\right)\\
&= \mathbb{P}\left(\left|Z^T\Sigma_0^{1/2}\mathrm{diag}(\widetilde{\Delta}_k))\Sigma_0^{1/2}Z - \mathbb{E}\left(Z^T\Sigma_0^{1/2}\mathrm{diag}(\widetilde{\Delta}_k)\Sigma_0^{1/2}Z\right)\right| \geq t_{\delta,k}\|\delta\|_2\right)\\
&\leq 2\exp\left[-c\min\left(\frac{t_k'^2}{\left\|\Sigma_0^{1/2}\mathrm{diag}(\widetilde{\Delta}_k)\Sigma_0^{1/2}\right\|_F^2}, \frac{t_k'}{\left\|\Sigma_0^{1/2}\mathrm{diag}(\widetilde{\Delta}_k)\Sigma_0^{1/2}\right\|_2}\right)\right] =: p_1,
\end{aligned}$$

where we use a shorthand notation

$$\left\|\Sigma_0^{1/2}\mathrm{diag}(\widetilde{\Delta}_k)\Sigma_0^{1/2}\right\|_F \leq \|\Sigma_0\|_2\left\|\mathrm{diag}(\widetilde{\Delta}_k)\right\|_F = m_k\|\Sigma_0\|_2$$

for $\delta \in \mathbb{S}^{d_k-1}$, since $\left\|\mathrm{diag}(\Delta'_{\Psi_k})\right\|_F = \|\delta\|_2 = 1$. Then for a fixed level of deviation $\delta \in \Pi_{d_k} \subset \mathbb{S}^{d_k-1}$, where Π_{d_k} is an ε-net of the sphere $\mathbb{S}^{d_k-1}$, we have $t_{\delta,k} \asymp \|\Sigma_0\|_2\left(\sqrt{p} \vee d_k\|\delta\|_\infty\right) = \|\Sigma_0\|_2\left(\sqrt{p} \vee d_k\left\|\mathrm{diag}(\Delta'_{\Psi_k})\right\|_2\right)$. Therefore, we have

$$\begin{aligned}
&\mathbb{P}\left(\exists\delta \in \Pi_{d_k} : \sum_{i=1}^{d_k}\delta_i\left(\langle Y_i^{(k)}, Y_i^{(k)}\rangle - \mathbb{E}\langle Y_i^{(k)}, Y_i^{(k)}\rangle\right) \geq t_{\delta,k}\right) =: \mathbb{P}(\text{ event } \mathcal{G}_k^c \text{ occurs })\\
&\leq (1+2/\varepsilon)^{d_k}p_1 \leq 2\exp\left(d_k\log 5 - c\min\left(\frac{p}{m_k}, \frac{d_k\left\|\mathrm{diag}(\Delta'_{\Psi_k})\right\|_2}{\left\|\mathrm{diag}(\Delta'_{\Psi_k})\right\|_2}\right)\right) \leq \exp(-cd_k\log 5).
\end{aligned}$$

Notice that the expression for $t_{k,\delta}$ clearly depends on the dimension d_k of Ψ_k. Moreover, the vector length $\|\delta\|_2$ is certainly affected by dimension d_k even when $\|\delta\|_\infty$ is at a fixed level across k. Then on event $\mathcal{G}_k$, for a chosen parameter $t_{\delta,k}$ as in (32), we have

$$\begin{aligned}
&\sup_{\delta\in\mathbb{S}^{d_k-1}}\sum_{i=1}^{d_k}\delta_i\left(\langle Y_i^{(k)}, Y_i^{(k)}\rangle - \mathbb{E}\langle Y_i^{(k)}, Y_i^{(k)}\rangle\right)\\
&\leq \frac{1}{1-\varepsilon}\sup_{\delta\in\Pi_{d_k}}\sum_{j=1}^{d_k}\delta_i\left(\langle Y_i^{(k)}, Y_i^{(k)}\rangle - \mathbb{E}\langle Y_i^{(k)}, Y_i^{(k)}\rangle\right) \leq \frac{t_{\delta,k}}{1-\varepsilon},
\end{aligned}$$

with a standard approximation argument as follows. Denote by

$$y = \left(\left\|Y_1^{(k)}\right\|_2^2 - \mathbb{E}\left\|Y_1^{(k)}\right\|_2^2, \ldots, \left\|Y_{d_k}^{(k)}\right\|_2^2 - \mathbb{E}\left\|Y_{d_k}^{(k)}\right\|_2^2\right).$$

We have for $\delta = (\delta_1, \ldots, \delta_{d_k})$

$$\sup_{\delta\in\Pi_{d_k}}\langle\delta, y\rangle \leq \|y\|_2 = \sup_{\delta\in\mathbb{S}^{d_k-1}}\langle y, \delta\rangle \leq \frac{1}{1-\varepsilon}\sup_{\delta\in\Pi_{d_k}}\langle\delta, y\rangle.$$

The LHS is obvious. To see the RHS, notice that for $\delta \in \mathbb{S}^{d_k-1}$ that achieves maximality in

$$\|y\|_2 = \sup_{\delta \in \mathbb{S}^{d_k-1}} \langle y, \delta \rangle ,$$

we can find $\delta_0 \in \Pi_{d_k}$ such that $\|\delta - \delta_0\|_2 \leq \varepsilon$. Now

$$\begin{aligned} \langle \delta_0, y \rangle &= \langle \delta, y \rangle - \langle \delta - \delta_0, y \rangle \\ &\geq \langle \delta, y \rangle - \sup_{\delta \in \mathbb{S}^{d_k-1}} \varepsilon \langle \delta, y \rangle = (1-\varepsilon) \sup_{\delta \in \mathbb{S}^{d_k-1}} \langle \delta, y \rangle , \\ \text{and hence} \quad \sup_{\delta \in \Pi_{d_k}} \langle \delta, y \rangle &\geq (1-\varepsilon) \sup_{\delta \in \mathbb{S}^{d_k-1}} \langle \delta, y \rangle = (1-\varepsilon) \|y\|_2 . \end{aligned}$$

The lemma thus holds. □

6. Experiments

In this section, we provide a suite of simulated data experiments confirming the tightness of our new rates. As we do not propose any changes to the TeraLasso estimator, we refer the reader to Greenewald et al. (2019a) for a detailed exploration of the application of TeraLasso to real world data sets.

Unless otherwise indicated, random graphs are created for each factor Ψ_k using an Erdos-Renyi (ER) topology, generated according to the method of Zhou et al. (2010). Initially we set $\Psi_k = 0.25 I_{n \times n}$, where $n = 100$, and randomly select q edges and update Ψ_k as follows: for each new edge (i, j), a weight $a > 0$ is chosen uniformly at random from $[0.2, 0.4]$; we subtract a from $[\Psi_k]_{ij}$ and $[\Psi_k]_{ji}$, and increase $[\Psi_k]_{ii}, [\Psi_k]_{jj}$ by a. This keeps Ψ_k (hence Ω_0) positive definite. We repeat this process until all edges are added. Finally, we form $\Omega_0 = \Psi_1 \oplus \cdots \oplus \Psi_K$. An example ER graph and precision matrix are shown in Figure 1.

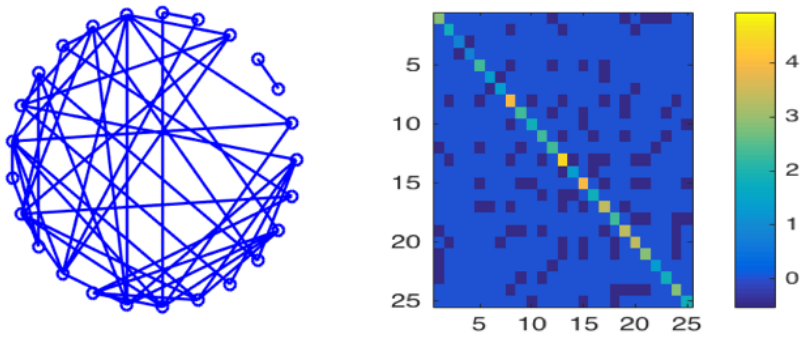


Figure 1: Example Erdos-Renyi random graph with 25 nodes and 50 edges. Left: Graphical representation. Right: Corresponding precision matrix Ψ.

The performance is empirically evaluated using several metrics including: the normalized Frobenius norm ($\|\widehat{\Omega} - \Omega_0\|_F / \|\Omega_0\|_F$) and spectral norm ($\|\widehat{\Omega} - \Omega_0\|_2$) error of the precision matrix estimate $\widehat{\Omega}$, and the precision/recall of edge recovery. We also use the Matthews correlation coefficient to quantify the edge misclassification error. Let the number of true

positive edge detections be TP, true negatives TN, false positives FP, and false negatives FN. The Matthews correlation coefficient is defined as Matthews (1975)

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}},$$

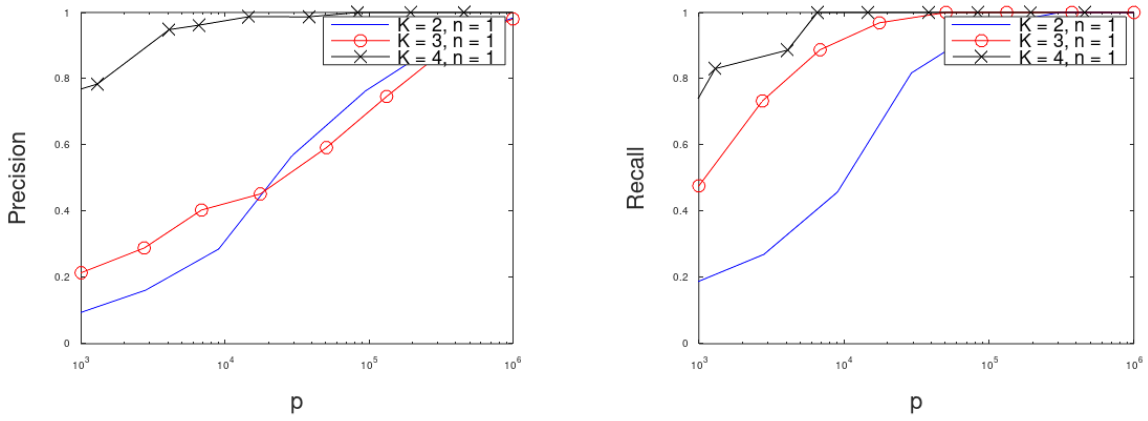
where each nonzero off diagonal element of Ψ_k is considered as a single edge. The MCC ranges from MCC = 0 implying complete failure to MCC = 1 implying perfect edge set estimation.

Throughout, the sparsity parameters ρ_k are set via the approach in the original TeraLasso paper Greenewald et al. (2019a) driven by the theoretical bounds therein. These are still valid in our present analysis since the improvement in the bounds focuses on the diagonal elements, not the off-diagonal ones. Figure 2 illustrates how increasing dimension p and K improves single sample performance, where the d_ks are not equal but proportional to each other. This plot is similar to the original plot Figure 6 in Greenewald et al. (2019a), where the d_ks are equal and scale together. Note that the precision and recall both still increase to 1 as p and K increase, though there is some performance degradation due to the non-matched d_k values, compared to the matching cases in Greenewald et al. (2019a). This degradation is as expected (since for the same p on the x axis, larger d_ks have smaller m_ks). Throughout this section, the results are averaged over 25 random trials.

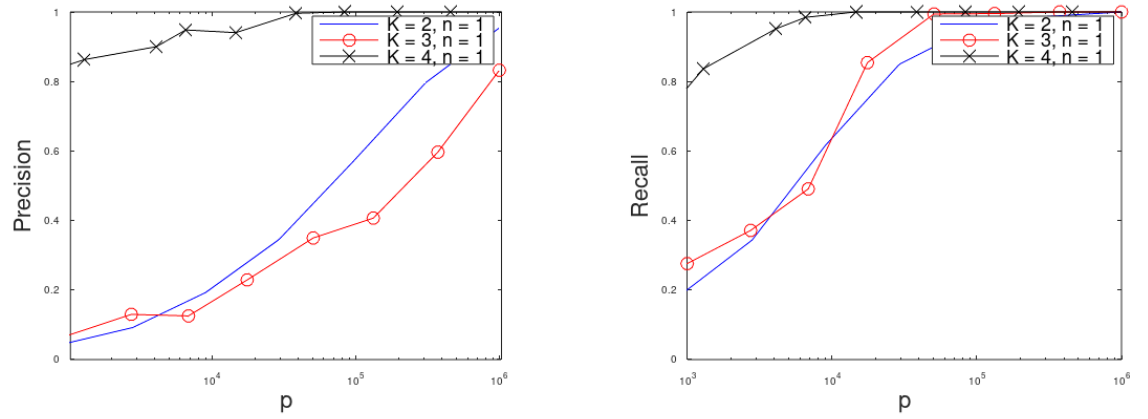
Chain graphs. Figure 3 shows results for $n = 1$ and the factors Ψ_k having sparsity patterns being chain graphs, i.e. the Ψ_k are tridiagonal (see the first row of Figure 3 for an graphical illustration and plot of Ψ_k for $d_k = 25$). In this potentially more challenging scenario, TeraLasso continues to succeed in estimation, with the increasing structure of higher K raising performance as predicted in Theorem 4. Note that Theorem 4 only guarantees $O(1)$ error for $K = 2$, yet we observe a small reduction in the spectral and normalized Frobenius errors as p increases even for $K = 2$. This is likely due to our bounds being worst-case (in particular Theorem 7 relating the spectral and Frobenius norms).

Varying sparsity and aspect ratio. In Figure 4, we fix $n = 10$, $K = 3$, and $p = 10^5$ and vary the ratio between the first factor dimension d_1 and the remaining factors $d_2 = d_3$ for various coefficients of $s_k \propto d_k$. Shown are the MCC, normalized Frobenius norm error, and normalized spectral norm error. Convergence is not expected as p and n are fixed. Note that as expected, increasing the graph densities (increasing s_k) degrades the performance, particularly the MCC. Furthermore, as d_1/p approaches 1 (so that d_1 becomes large and m_1 becomes small) the error becomes large. This confirms the theoretical results that generally say error scales with increasing $d_{\max}/m_{\min}$.

Tightness of the spectral norm error bounds. We next conduct a $K = 3$ experiment to confirm the tightness of the $n = 1$ spectral norm error bounds as they relate to the aspect ratio in Corollary 6. For each p, we set $d_1 = \sqrt{p}$ and $d_2 = d_3 = p^{1/4}$, and set $s_k = \frac{d_k}{\log_{10} p}$ (rounding in each case to the nearest integer), which ensures that the $\frac{s_k \log p}{d_k}$ terms in Corollary 6 are $O(1)$. This corresponds to an overall $s = \sum_{k=1}^{K} s_k m_k = \frac{p}{\log_{10} p}$, which implies that Assumption (A3) with $n = 1$ requires $(m_{\min})^2 = \Omega(p)$, which is tightly satisfied since here $m_{\min} = \sqrt{p}$. Applying Corollary 6, we have that the spectral norm error will scale as $O_p\left(\sqrt{\frac{d_{\max}}{m_{\min}}}\right) = O_p\left(\sqrt{\frac{d_1}{m_1}}\right) = O_p(1)$, hence the spectral norm error bound does not converge. The experimental results are shown in Figure 5, where indeed the spectral



(a) Rectangular with scaling 2. $K = 2$: $d_2 = 2d_1$. $K = 3$: $\frac{1}{2}d_2 = d_1 = 2d_3$. $K = 4$: $\frac{1}{2}d_4 = \frac{1}{2}d_2 = d_1 = 2d_3$.



(b) Rectangular with scaling 4. $K = 2$: $d_2 = 4d_1$. $K = 3$: $\frac{1}{4}d_2 = d_1 = 4d_3$. $K = 4$: $\frac{1}{4}d_4 = \frac{1}{4}d_2 = d_1 = 4d_3$.

Figure 2: Edge support estimation on random ER graphs, with the ρ_k set according to A1. Graphical model edge detection precision and recall curves are shown as a function of data dimension $p = \prod_{k=1}^{K} d_k$. Shown here are non-rectangular edge support estimation on random ER graphs. For each value of the tensor order K, we set the $d_k \propto p^{1/K}$ with the constants of proportionality shown in the captions. Observe single sample convergence as the dimension p increases and as increasing K creates additional structure.

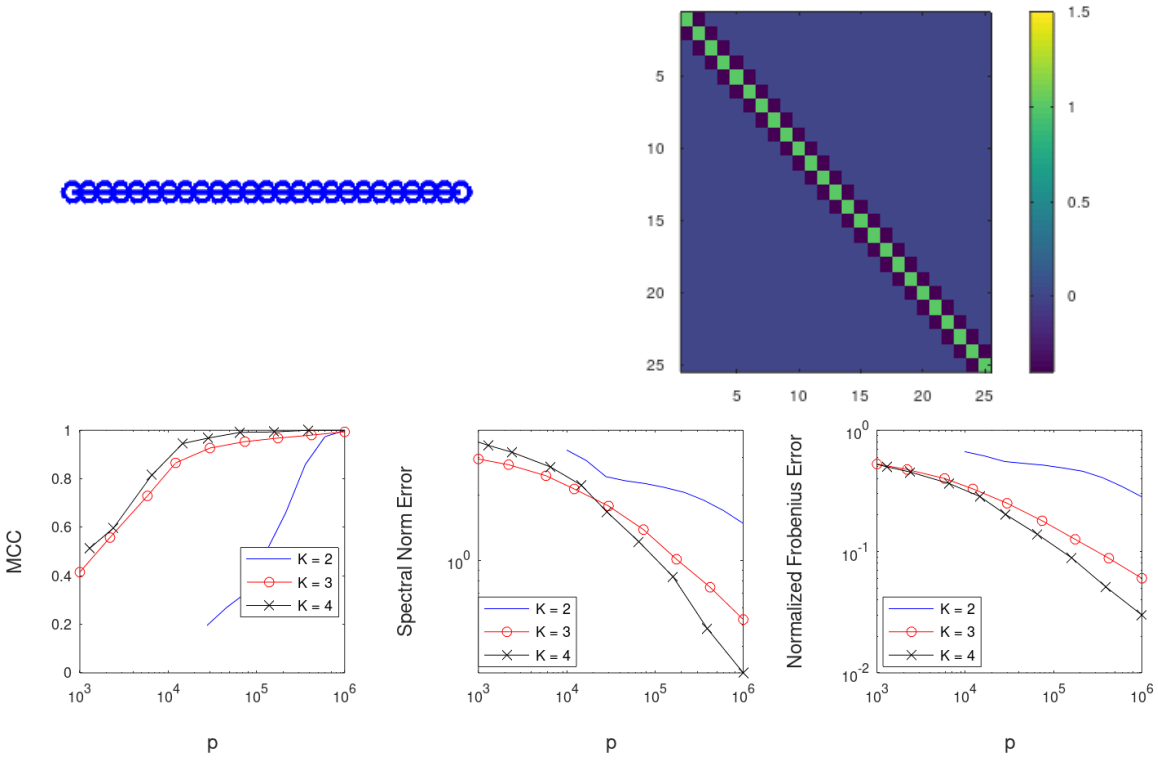


Figure 3: Edge support estimation for Ψ_k deterministic chain graphs. Top row: Example graphical representation of Ψ_k structure and image of the corresponding Ψ_k for $d_k = 25$. Bottom row: Graphical model edge detection precision and recall curves are shown as a function of data dimension $p = \prod_{k=1}^{K} d_k$. For each value of the tensor order K, we set $d_k = p^{1/K}$. Observe single sample convergence as the dimension p increases and as increasing K creates additional structure.

norm error is found to not converge and to stay roughly constant consistent with the $O_p(1)$ bound. This indicates that the spectral norm error bound does seem to be tight, i.e. it cannot be improved for this configuration of parameters. On the other hand, both the precision and recall converge, indicating that the support set of the sparse precision matrix entries is still being successfully recovered. This is possible because the spectral norm error is a more stringent measure of performance as it considers errors in all possible linear projections of the precision matrix.

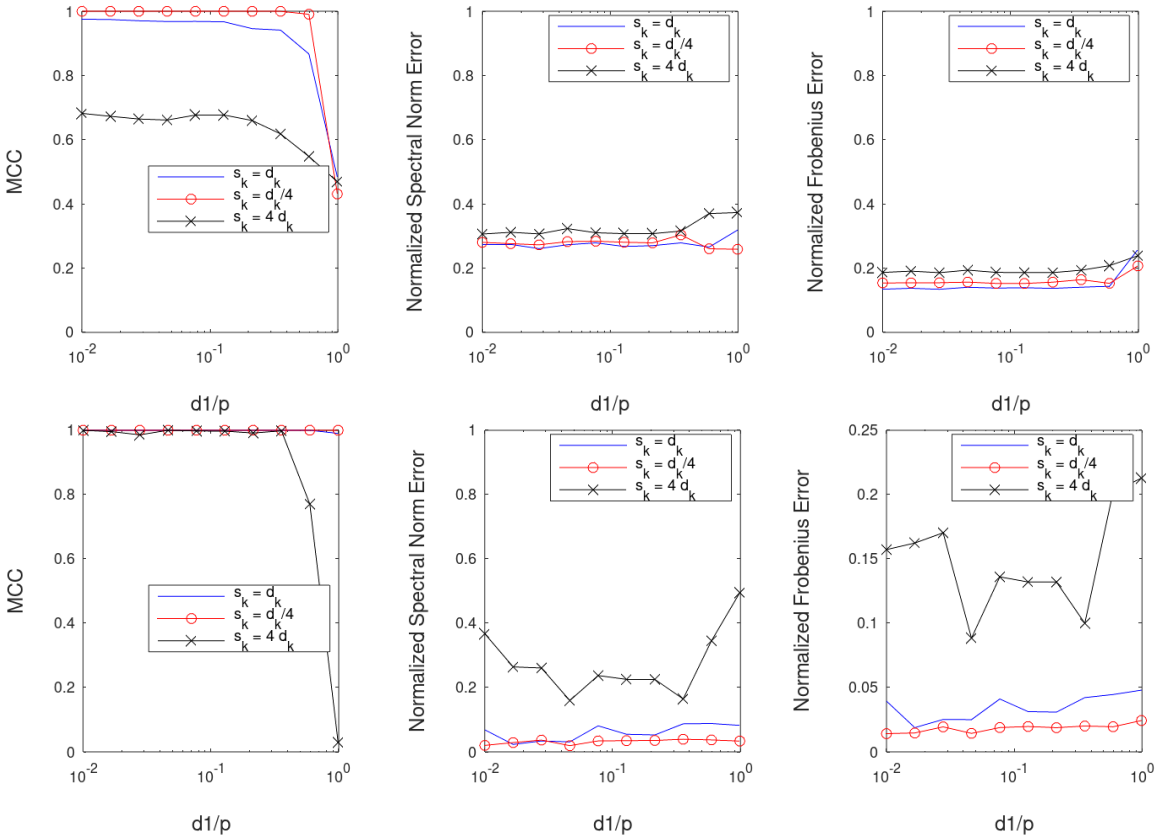


Figure 4: MCC, Frobenius norm error, and spectral norm error for varying relative dimensions (d_k) for $n = 10$ (top) and $n = 1000$ (bottom), $K = 3$ and $p = 10^5$: d_1 set according to Aspect Ratio on the x axis, $d_2 = d_3 = \sqrt{p/d_1}$. Results are shown for different sparsity levels of the random ER graphs for each factor. Note that performance is relatively stable across aspect ratios, until d_1/p becomes too large such that d_1 vastly dominates $m_1 = d_2 d_3$, degrading performance. Additionally, increasing sparsity decreases performance as the problem becomes more difficult.

7. Conclusion

In this work, we present sharp rates of convergence for the ℓ_1-regularized TeraLasso estimator of precision matrices with Kronecker sum structures. The improved bounds critically expand the settings for which finite sample ($n = O(1)$) error bounds can be derived from the $K \geq 3$ regime of Greenewald et al. (2019a) to now include the $K = 2$ regime. This closes the gap between the previous theoretical bounds and the empirical observation of $K = 2$ convergence. This guarantee now justifies for principled application of the TeraLasso estimator to the large body of problems for which two-way structure applies, e.g. spatio-temporal and matrix data.

Acknowledgments and Disclosure of Funding

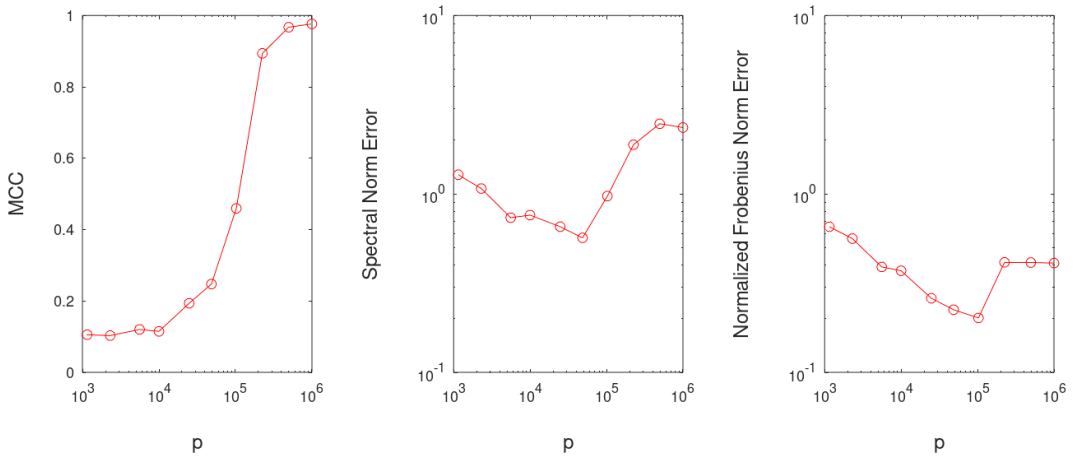


Figure 5: Verifying tightness of the spectral norm error bound: for $K = 3$ and $d_1 = \sqrt{p}$ and $d_2 = d_3 = p^{1/4}$, Corollary 6 predicts $O_p(1)$ spectral norm error, which we observe in this experiment (Erdos-Renyi graphs, $s_k = \frac{d_k}{\log_{10} p}$).

We thank Harrison Zhou for helpful discussions. We thank the Simons Institute for the Theory of Computing at University of California, Berkeley at the occasion of Algorithmic Advances for Statistical Inference with Combinatorial Structure Workshop, and organizers of International Conference on Statistics and Related Fields (ICON STARF), University of Luxembourg, for their kind invitations, where this work was presented in part.

A. Proof of preliminary results in Section 4

A.1 Proof of Lemma 12

First, we prove (26). We have by (25)

$$\begin{aligned}
&\Delta_g + \langle \operatorname{offd}(\Delta_\Omega), \widehat{S} - \Sigma_0 \rangle \\
&\quad \geq \sum_{k=1}^{K} m_k \rho_k \left(|\Psi_k + \Delta_{\Psi_k}|_{1,\text{off}} - |\Psi_k|_{1,\text{off}} \right) + \langle \operatorname{offd}(\Delta_\Omega), \widehat{S} - \Sigma_0 \rangle =: S_2
\end{aligned}$$

where under the settings of Lemma 10,

$$\begin{aligned}
S_2 &\geq \sum_{k=1}^{K} m_k \rho_k \left(|\Delta_{\Psi_k, S^c}|_1 - |\Delta_{\Psi_k, S}|_1 \right) - \sum_{k=1}^{K} m_k |\Delta_{\Psi_k}|_{1,\text{off}} \delta_k \\
&\geq \sum_{k=1}^{K} m_k \rho_k \left(|\Delta_{\Psi_k, S^c}|_1 - |\Delta_{\Psi_k, S}|_1 \right) - \sum_{k=1}^{K} m_k \delta_k \left(|\Delta_{\Psi_k, S^c}|_1 + |\Delta_{\Psi_k, S}|_1 \right) \\
&\geq -\sum_{k=1}^{K} m_k (\rho_k + \delta_k) |\Delta_{\Psi_k, S}|_1 \geq -2 \max_k \rho_k \sum_{k=1}^{K} m_k |\Delta_{\Psi_k, S}|_1 \\
&= -2 \max_k \rho_k |\Delta_S|_1 ;
\end{aligned}$$

Thus (26) holds. □

A.2 Proof of Lemma 17

We first state Proposition 24

Proposition 24 *Under (A1)-(A3), for all* $\Delta \in \mathcal{T}_n$,

$$\|\Delta\|_2 \leq M r_{\mathbf{p}} \sqrt{\frac{K+1}{\min_k m_k}} \leq \frac{1}{2}\phi_{\min}(\Omega_0) \tag{43}$$

so that $\Omega_0 + v\Delta \succ 0, \forall v \in I \supset [0,1]$, *where* I *is an open interval containing* $[0,1]$.

Proof By Proposition 14, (43) holds for $\Delta \in \mathcal{T}_n$; Next, it is sufficient to show that $\Omega_0 + (1+\varepsilon)\Delta \succ 0$ and $\Omega_0 - \varepsilon\Delta \succ 0$ for some $1 > \varepsilon > 0$. Indeed, for $\varepsilon < 1$,

$$\begin{aligned} \phi_{\min}(\Omega_0 + (1+\varepsilon)\Delta) &\geq \phi_{\min}(\Omega_0) - (1+\varepsilon)\|\Delta\|_2 \\ &> \phi_{\min}(\Omega_0) - 2\sqrt{\frac{K+1}{\min_k m_k}} M r_{\mathbf{p}} > 0 \end{aligned}$$

given that by definition of $\mathcal{T}_n$ and (43). ■

Thus we have that $\log|\Omega_0 + v\Delta|$ is infinitely differentiable on the open interval $I \supset [0,1]$ of v. This allows us to use the Taylor's formula with integral remainder to prove Lemma 17, drawn from Rothman et al. (2008),

Let us use A as a shorthand for

$$\mathrm{vec}\{\,\Delta\,\}^{\mathrm{T}} \left(\int_0^1 (1-\mathrm{v})(\Omega_0 + \mathrm{v}\Delta)^{-1} \otimes (\Omega_0 + \mathrm{v}\Delta)^{-1} \mathrm{dv} \right) \mathrm{vec}\{\,\Delta\,\},$$

where $\mathrm{vec}\{\,\Delta\,\} \in \mathbb{R}^{\mathrm{p}^2}$ is $\Delta_{p\times p}$ vectorized. Now, the Taylor expansion gives

$$\begin{aligned} \log|\Omega_0 + \Delta| - \log|\Omega_0| &= \frac{d}{dv}\log|\Omega_0 + v\Delta|\Big|_{v=0} \Delta \\ &\quad + \int_0^1 (1-v)\frac{d^2}{dv^2}\log|\Omega_0 + v\Delta|\,dv = \langle\,\Sigma_0, \Delta\,\rangle - A; \end{aligned} \tag{44}$$

The last inequality holds because $\nabla_\Omega \log|\Omega| = \Omega^{-1}$ and $\Omega_0^{-1} = \Sigma_0$. We now bound A, following arguments from Zhou et al. (2011); Rothman et al. (2008)

$$\begin{aligned} A &= \int_0^1 (1-v)\frac{d^2}{dv^2}\log|\Omega_0 + v\Delta| dv \\ &= \mathrm{vec}(\Delta)^T \left(\int_0^1 (1-v)(\Omega_0 + v\Delta)^{-1} \otimes (\Omega_0 + v\Delta)^{-1} dv) \right) \mathrm{vec}(\Delta) \\ &\geq \|\Delta\|_F^2 \phi_{\min} \left(\int_0^1 (1-v)(\Omega_0 + v\Delta)^{-1} \otimes (\Omega_0 + v\Delta)^{-1} dv \right). \end{aligned}$$

Now,

$$\begin{aligned}
&\phi_{\min}\left(\int_0^1 (1-v)(\Omega_0+v\Delta)^{-1}\otimes(\Omega_0+v\Delta)^{-1}dv\right)\\
&\geq \int_0^1 (1-v)\phi^2_{\min}((\Omega_0+v\Delta)^{-1})dv\\
&\geq \min_{v\in[0,1]}\phi^2_{\min}((\Omega_0+v\Delta)^{-1})\int_0^1(1-v)dv\\
&= \frac{1}{2}\min_{v\in[0,1]}\frac{1}{\phi^2_{\max}(\Omega_0+v\Delta)} = \frac{1}{2\max_{v\in[0,1]}\phi^2_{\max}(\Omega_0+v\Delta)}\\
&\geq \frac{1}{2(\phi_{\max}(\Omega_0)+\|\Delta\|_2)^2}.
\end{aligned}$$

where by (43), we have for all $\Delta\in\mathcal{T}_n$,

$$\|\Delta\|_2 \leq \sqrt{\frac{K+1}{\min_k m_k}}\|\Delta\|_F = \sqrt{\frac{K+1}{\min_k m_k}}Mr_{\mathbf{p}} < \frac{1}{2}\phi_{\min}(\Omega_0)$$

so long as the condition in (A3) holds, namely,

$$(\min_k m_k)^2 > 2C^2\kappa(\Sigma_0)^4(s\log p + Kp)(K+1)$$

Hence,

$$\phi_{\min}\left(\int_0^1 (1-v)(\Omega_0+v\Delta)^{-1}\otimes(\Omega_0+v\Delta)^{-1}dv\right) \geq \frac{2}{9\phi^2_{\max}(\Omega_0)}.$$

Thus, substituting into (44), the lemma is proved. □

B. Proof sketch Theorem 5

First, we clarify the trace identifiability conditions. Then, we state the key concentration events whose probability we control, which are crucial to prove Theorem 4 and Lemma 2.

B.1 Factor identifiability up to traces

Observe that the set $\mathcal{K}_{\mathbf{p}}$ (5) of matrices expressible as a Kronecker sum is a linear sum of K components, and thus $\mathcal{K}_{\mathbf{p}}$ is linearly spanned by the K components. Thus $\mathcal{K}_{\mathbf{p}}$ is a linear subspace of $\mathbb{R}^{p\times p}$. Note that $\Omega = \Psi_1\oplus\cdots\oplus\Psi_K$ does not uniquely determine $\{\Psi_k\}_{k=1}^K$, i.e., without further constraints the Kronecker sum parameterization is not fully identifiable. Specifically, observe that for any c

$$A\oplus B = A\otimes I + I\otimes B = A\otimes I - cI + cI + I\otimes B = (A-cI)\oplus(B+cI),$$

and thus the trace of each factor is non-identifiable. More generally,

$$\Omega = (\Psi_1 + c_1 I_{d_1})\oplus\cdots\oplus(\Psi_K + c_K I_{d_K}) = \Psi_1\oplus\cdots\oplus\Psi_K, \tag{45}$$

whenever c_k are any scalars such that $\sum_{k=1}^{K} c_k = 0$. Observe also that $\mathrm{offd}(\Psi_k)$, $\mathrm{diag}(\Omega)$ are unaffected by this trace ambiguity, where recall $\mathrm{offd}(M) = M - \mathrm{diag}(M)$. Thus, this trace ambiguity does not affect Ω as a whole or the off diagonals of the factors Ψ_k. Moreover, the decomposition

$$\Omega = \mathrm{diag}(\Omega) + \mathrm{offd}(\Psi_1 \oplus \cdots \oplus \Psi_K) \tag{46}$$

is thus identifiable. Hence the parameter $\tau_\Omega := \mathrm{tr}(\Omega)/p$ is uniquely determined. We note that, as a result, the "trace non-identifiability" cannot impact the interpretation of the estimated model, and hence it is sufficient to adopt a convention by requiring

$$\mathrm{trace}(\Psi_1)/d_1 = \cdots = \mathrm{trace}(\Psi_K)/d_K, \quad \text{in view of} \tag{45}$$

on the traces of Ψ_k to make all parameters identifiable. In particular, we use the convention as specified in Lemma 21 when we deal with $\Delta_\Omega := \Omega - \Omega_0$ for the rest of the paper. For more details about an identifiable parameterization and other conventions, see Greenewald et al. (2019a).

B.2 Definition of existing key concentration events

Our previous work in Greenewald et al. (2019a) relied on component-wise concentration bounds for sample covariance S_n^k around its mean $\Sigma_0^{(k)}, k \in [K]$. In words, $\mathcal{T}$ is the event that $\delta_{n,k}$ dominates the maximum of entrywise errors for estimating the off-diagonal elements of covariance matrix $\Sigma^{(k)}$ as in (8) using S_n^k, simultaneously for all k,

$$\left\| \mathrm{offd}(S_n^k) - \mathrm{offd}(\Sigma_0^{(k)}) \right\|_{\max} \leq \delta_{n,k} \quad \forall k; \tag{47}$$

The same upper bound also holds for the diagonal component, cf. Lemma 25, leading to the error bound on the diagonal component of the loss function as stated in Lemma 13 Greenewald et al. (2019a). In fact, the off-diagonal statement in the following Lemma 25 is a generalization of Lemma 10, where $n = 1$, to sample size $n \geq 1$, which follows from Lemma 12 Greenewald et al. (2019b). For the diagonal elements, we use a set of new events as defined in Lemma 20 in the present work. To set up the discussion, we also present the original concentration of measure bound, which follows from Lemma 13 Greenewald et al. (2019b). More precisely, we have Lemma 25. These error probabilities are then combined via the union bound.

Lemma 25 Greenewald et al. (2019a) *On event $\mathcal{T}$, we have* (24) *holds for*

$$\delta_{n,k} \asymp \|\Sigma_0\|_2 \sqrt{\frac{\log p}{n m_k}}$$

as in (10)*. On event $\mathcal{D}_0$ and the following event $\mathcal{T}'$*

$$\textit{event } \mathcal{T}' \;:=\; \left\{ \left\| \mathrm{diag}(S_n^k) - \mathrm{diag}(\Sigma_0^{(k)}) \right\|_{\max} \leq \delta_{n,k} \quad \forall k \right\}, \tag{48}$$

we have for all $\Delta_\Omega \in \mathcal{K}_{\mathbf{p}}$,

$$\frac{\left| \langle\, \mathrm{diag}(\Delta_\Omega), \widehat{S} - \Sigma_0 \,\rangle \right|}{\sqrt{K+1}\, \|\mathrm{diag}(\Delta_\Omega)\|_F} \;\leq\; C_1 \max_k \delta_{n,k} \sqrt{p} \asymp \|\Sigma_0\|_2 \max_k \sqrt{\frac{d_k \log p}{n}}.$$

Moreover, $\mathbb{P}\left(\mathcal{T} \cap \mathcal{T}' \cap \mathcal{D}_0\right) \geq 1 - 2K \exp(-c \log p)$ for some absolute constants c, C_1.

Finally, under the joint event $\mathcal{T} \cap \mathcal{T}' \cap \mathcal{D}_0$, the error in the Frobenius norm is bounded in the sense of Theorem 5, following the sequence of arguments as in Section 4.

References

G. Allen and R. Tibshirani. Transposable regularized covariance models with an application to missing data imputation. *Ann. Appl. Stat.*, 4(2):764–790, 2010.

S. Andrianov. A matrix representation of lie algebraic methods for design of nonlinear beam lines. In *AIP Conf. Proc.*, volume 391, pages 355–360. AIP, 1997.

O. Banerjee, L. E. Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *JMLR*, 9:485–516, 2008.

B. Beckermann, D. Kressner, and C. Tobler. An error analysis of Galerkin projection methods for linear systems with tensor product structure. *SINUM*, 51(6):3307–3326, 2013.

A. Chapman, M. Nabi-Abdolyousefi, and M. Mesbahi. Controllability and observability of network-of-networks via cartesian products. *IEEE Trans. on Automatic Control*, 59(10): 2668–2679, 2014.

A. d'Aspremont, O. Banerjee, and L. E. Ghaoui. First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 30:56–66, 2008.

A. P. Dawid. Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika*, 68:265–274, 1981.

F. Dorr. The direct solution of the discrete Poisson equation on a rectangle. *SIAM review*, 12(2):248–263, 1970.

P. Eilers and B. Marx. Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and intelligent laboratory sys.*, 66 (2):159–174, 2003.

N. Ellner. New ADI model problem applications. In *Proc. of ACM Fall joint computer conf.*, pages 528–534, 1986.

M. Fey, J. Eric Lenssen, F. Weichert, and H. Müller. Splinecnn: Fast geometric deep learning with continuous b-spline kernels. In *CVPR*, pages 869–877, 2018.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, 9(3):432–441, December 2008.

L. Grasedyck. Existence and computation of low Kronecker-rank approximations for large linear systems of tensor product structure. *Computing*, 72(3-4):247–265, 2004.

K. Greenewald, S. Zhou, and A. Hero. The Tensor graphical Lasso (TeraLasso). *J. R. Stat. Soc., B: Stat. Methodol.*, 81(5):901–931, 2019a.

K. Greenewald, S. Zhou, and A. Hero. Supplementary material for "Tensor Graphical Lasso (teralasso)". *J. R. Stat. Soc., B: Stat. Methodol.*, 2019b.

A. Gupta and T. Varga. Characterization of matrix variate normal distributions. *J. Multivar. Anal.*, 41:80–88, 1992.

M. Hornstein, R. Fan, K. Shedden, and S. Zhou. Joint mean and covariance estimation for unreplicated matrix-variate data. *J. Amer. Statist. Assoc.*, 114(526):682–696, 2019.

W. Imrich, S. Klavžar, and D. F. Rall. *Topics in graph theory: Graphs and their Cartesian product*. AK Peters/CRC Press, 2008.

A. Kalaitzis, J. Lafferty, N. Lawrence, and S. Zhou. The Bigraphical Lasso. In *ICML*, pages 1229–1237, 2013.

T. Kolda and B. Bader. Tensor decompositions and applications. *SIAM review*, 51(3): 455–500, 2009.

M. Kotzagiannidis and P. Dragotti. Splines and wavelets on circulant graphs. *Appl. and Comp. Harmonic Anal.*, 2017.

D. Kressner and C. Tobler. Krylov subspace methods for linear systems with tensor product structure. *SIMAX*, 31(4):1688–1714, 2010.

S. Lauritzen. *Graphical Models*. Oxford Univ. Press, 1996.

C. Leng and C. Tang. Sparse matrix graphical models. *J. Amer. Statist. Assoc.*, 107: 1187–1200, 2012.

S. Li, M. López-Garcia, N. D. Lawrence, and L. Cutillo. Two-way sparse network inference for count data. In *AISTATS*, pages 10924–10938. PMLR, 2022.

D. Luenberger. Observers for multivariable systems. *IEEE Trans. on Automatic Control*, 11(2):190–197, 1966.

B. W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta-Protein Structure*, 405(2):442–451, 1975.

N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the Lasso. *Ann. Statist.*, 34(3):1436–1462, 2006.

V. D. Milman and G. Schechtman. *Asymptotic Theory of Finite Dimensional Normed Spaces. Lecture Notes in Mathematics 1200*. Springer, 1986.

A. Rothman, P. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electron. J. Stat.*, 2:494–515, 2008.

M. Rudelson and R. Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electron. Commun. Probab.*, 18:1–9, 2013.

M. Rudelson and S. Zhou. Errors-in-variables models with dependent measurements. *Electron. J. Statist.*, 11(1):1699–1797, 2017.

U. Schmitt, A. K. Louis, F. Darvas, H. Buchner, and M. Fuchs. Numerical aspects of spatio-temporal current density reconstruction from EEG-/MEG-data. *IEEE Trans. on Medical Imaging*, 20(4):314–324, 2001.

X. Shi, Y. Wei, and S. Ling. Backward error and perturbation bounds for high order sylvester tensor equation. *Lin. and Multilin. Alg.*, 61(10):1436–1446, 2013.

T. Tsiligkaridis, A. Hero, and S. Zhou. On convergence of kronecker graphical lasso algorithms. *IEEE Trans. Signal Process.*, 61:1743–1755, 2013.

C. F. Van Loan. The ubiquitous Kronecker product. *Journal of Comp. and Appl. Math.*, 123(1-2):85–100, 2000.

Y. Wang and A. Hero. SG-PALM: a fast physically interpretable tensor graphical model. *arXiv preprint arXiv:2105.12271*, 2021.

Y. Wang, Z. Sun, D. Song, and A. Hero. Kronecker-structured covariance models for multiway data. *arXiv preprint arXiv:2212.01721*, 2022.

S. Wood, N. Pya, and B. Safken. Smoothing parameter and model selection for general smooth models. *Journal of the Amer. Stat. Assoc.*, 111(516):1548–1563, 2016.

S. N. Wood. Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, 62(4):1025–1036, 2006.

J. H. Yoon and S. Kim. EiGLasso for scalable sparse Kronecker-sum inverse covariance estimation. *Journal of Machine Learning Research*, 23, 2022.

M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94:19–35, 2007.

S. Zhou. Gemini: Graph estimation with matrix variate normal instances. *Ann. Statist.*, 42(2):532–562, 2014.

S. Zhou. Concentration of measure bounds for matrix-variate data with missing values. *Bernoulli*, 2023. to appear.

S. Zhou, J. Lafferty, and L. Wasserman. Time varying undirected graphs. *Mach. Learn.*, 80(2–3):295–319, September 2010.

S. Zhou, P. Rütimann, M. Xu, and P. Bühlmann. High-dimensional covariance estimation based on Gaussian graphical models. *JMLR*, 12:2975–3026, 2011.