Supplementary Material for Tensor Graphical Lasso (TeraLasso)

1. Supplement outline

This supplement is organized as follows. Sections 2-3 focus on the implementation and numerical convergence of the TeraLasso algorithm and Sections 4-8 focus on theory and proofs of convergence. Section 2 presents the algorithm for TeraLasso with nonconvex regularization and describes additional properties of the TeraLasso algorithm, including a discussion of the choice of step size, decomposition of the gradient update, and proof of joint convexity of the objective. Section 3 presents additional numerical experiments, including convergence of the nonconvex algorithm, larger scale TG-ISTA convergence experiments, additional discussion comparing the fit of the TeraLasso model to the wind speed data, and a discussion of the geometric differences between the Gemini and TeraLasso objectives.

We then proceed to the convergence analysis. Section 4 describes properties of the Kronecker sum and the Kronecker sum subspace $\mathcal{K}_{\mathbf{p}}$ that are needed for the remainder of the discussion. Proof of the main Frobenius norm theorem and of the spectral norm theorem are in Section 5, with the concentration bounds proven in Section 6. Section 7 proves the result on nonconvex regularization, and Section 8 presents and proves theorems on the geometrical convergence of the TG-ISTA algorithm. Relevant properties and identities relating to the space $\mathcal{K}_{\mathbf{b}}$ spanned by Kronecker sum matrices are contained in Appendix A, and a discussion of the case where the diagonal elements of Ω are known is given in Appendix B.

2. TeraLasso algorithm step size and numerical convergence proofs

2.1. Convergence of nonconvex regularization algorithm

The TG-ISTA implementation of the TeraLasso algorithm for nonconvex regularizers is shown in Algorithm 2. The primary differences from the $\ell 1$ regularized case are (a) the addition of the norm constraint, and (b) the use of the nonconvex regularizer in the gradient computation.

Algorithm 2 TG-ISTA implementation of TeraLasso with nonconvex regularization

```
Input: SCM factors $S_k$, regularization parameter $\rho$, regularizer $g_\rho(\cdot)$ and associated $q'_\rho(\cdot)$, backtracking constant $c \in (0,1)$, initial step size $\zeta_{1,0}$, initial iterate $\Omega_{\mathrm{init}} = I \in \mathcal{K}^\sharp_{\mathbf{p}}$.
while not converged do
    Compute the subspace gradient $\mathrm{Proj}_{\mathcal{K}_{\mathbf{p}}}\left(\Omega_t^{-1}\right) = G_1^t \oplus \cdots \oplus G_K^t$.
    Line search: Let stepsize $\zeta_t$ be the largest element of $\{c^j \zeta_{t,0}\}_{j=1,\ldots}$ such that the following are satisfied for $\Psi_k^{t+1} = \mathrm{shrink}^-_{\zeta_t \rho}(\Psi_k^t - \zeta_t(\widetilde{S}_k - G_k^t + q'_\rho(\Psi_k)))$:
        (a) $\|\Psi_1^{t+1} \oplus \cdots \oplus \Psi_K^{t+1}\|_2 \le \kappa$,
        (b) $\Psi_1^{t+1} \oplus \cdots \oplus \Psi_K^{t+1} \succ 0$,
        (c) $f(\{\Psi_k^{t+1}\}) \le \mathcal{Q}_{\zeta_t}(\{\Psi_k^{t+1}\}, \{\Psi_k^{t+1}\})$.
    for $k = 1, \ldots, K$ do
        Composite objective gradient update:
            $$\Psi_k^{t+1} \leftarrow \mathrm{shrink}^-_{\zeta_t \rho}\left(\Psi_k^t - \zeta_t(\widetilde{S}_k - G_k^t + q'_\rho(\Psi_k))\right).$$
    end for
    Compute next Barzilai-Borwein stepsize $\zeta_{t+1,0}$ via (27) in supplement 2.2.
end while
Return $\{\Psi_k^{t+1}\}_{k=1}^K$.
```

2.2. Choice of step size ζ_t

Here we propose a method (25) for selecting the stepsize parameter ζ_t at each step t that ensures convergence of the algorithm. We follow the approach of Beck and Teboulle (2009) and Guillot et al. (2012). Since $\Omega_t \succ 0$ and the the positive definite cone is an open set, there will always exist a ζ_t small enough such that $\Omega_{t+1} \succ 0$. We prove geometric convergence when ζ_t is chosen such that $\Omega_{t+1} \succ 0$ and

$$f(\Omega_{t+1}) = -\log|\Omega_{t+1}| + \langle \widehat{S}, \Omega_{t+1} \rangle \le \mathcal{Q}_{\zeta_t}(\Omega_{t+1}, \Omega_t) \tag{25}$$

where $\mathcal{Q}_{\zeta_t}$ is a quadratic approximation to f given by

$$\begin{aligned} &\mathcal{Q}_{\zeta_t}(\Omega_{t+1}, \Omega_t) \\ &\quad = -\log|\Omega_t| + \langle \widehat{S}, \Omega_t \rangle + \langle \Omega_{t+1} - \Omega_t, \nabla f(\Omega_t) \rangle + \frac{1}{2\zeta_t}\|\Omega_{t+1} - \Omega_t\|_F^2. \end{aligned} \tag{26}$$

At each iteration t, we thus perform a line search to select an appropriate ζ_t. We first select an initial stepsize $\zeta_{t,0}$ and compute the update (19). If the resulting Ω_{t+1} is not positive definite or does not decrease the objective sufficiently according to (25), we decrease the stepsize ζ_t to $c\zeta_{t,0}$ for $c \in (0,1)$ and re-evaluate if the resulting Ω_{t+1} satisfies the conditions. This backtracking process is repeated (setting stepsize equal to $c^j\zeta_{t,0}$ where j is incremented) until the resulting Ω_{t+1} satisfies the conditions. Since by construction Ω_t is positive definite, and the positive definite cone is an open set, there will be a step size small enough such that the conditions are satisfied. In practice, if after a set number of backtracking steps the conditions are still not satisfied, we can always take the safe step

$$\zeta_t = \lambda^2_{\min}(\Omega_t) = \sum_{k=1}^{K} \min_i [\mathbf{s}_k]_i^2.$$

As the safe stepsize often leads to slower convergence, we use the more aggressive Barzilai-Borwein step to set a starting $\zeta_{t,0}$ at each time. The Barzilai-Borwein stepsize presented in Barzilai and Borwein (1988) creates an approximation to the Hessian, in our case given by

$$\zeta_{t+1,0} = \frac{\|\Omega_{t+1} - \Omega_t\|_F^2}{\langle \Omega_{t+1} - \Omega_t, \nabla f(\Omega_t) - \nabla f(\Omega_{t+1})\rangle} \tag{27}$$

We derive the gradient $\nabla f(\Omega_t)$ in the next section. The norms and inner products in (27) and (26) can be efficiently computed factorwise (using the Ψ_k and S_k only) using the formulas in Appendix A.1.

2.3. Generation of Kronecker Sum Random Tensors

Generating random tensors given a Kronecker sum precision matrix can be made efficient by exploiting the Kronecker sum eigenstructure. Algorithm 3 allows efficient generation of data following the TeraLasso model.

2.4. Detailed TeraLasso Algorithm

Algorithm 4 shows additional details of the implementation of Algorithm 1 in the main text.

2.5. Decomposition of Objective: Proof of Lemma 4

For simplicity of notation define G_t to be the projection of Ω^{-1} onto the cone $\mathcal{K}_{\mathbf{p}}$ of positive definite Kronecker sum matrices:

$$G_t = G_1^t \oplus \cdots \oplus G_K^t = \mathrm{Proj}_{\mathcal{K}_{\mathbf{p}}}(\Omega_t^{-1}).$$

Algorithm 3 Generation of subgaussian tensor $X \in \mathbb{R}^{d_1\times\cdots\times d_K}$ under TeraLasso model.

```
Assume $\Sigma^{-1} = \Psi_1 \oplus \cdots \oplus \Psi_K$.
Input precision matrix factors $\Psi_k \in \mathbb{R}^{d_k\times d_k}$, $k = 1, \dots, K$.
**for** $k = 1, \dots, K$ **do**
    $U_k, \Lambda_k \leftarrow \mathrm{EIG}(\Psi_k)$ eigendecomposition of $\Psi_k$.
**end for**
$\mathbf{v} = [v_1, \dots, v_p] \leftarrow \mathrm{diag}(\Lambda_1) \oplus \cdots \oplus \mathrm{diag}(\Lambda_K) \in \mathbb{R}^p$.
Generate isotropic subgaussian random vector $z \in \mathbb{R}^p$.
$\widetilde{x}_i \leftarrow v_i^{-1/2} z_i$, for $i = 1, \dots, p$.
**for** $k = 1, \dots, K$ **do**
    $\widetilde{\mathbf{x}} \leftarrow (I_{[d_{1:k-1}]} \otimes U_k \otimes I_{[d_{k+1:K}]})\widetilde{\mathbf{x}}$.
**end for**
Reshape $\widetilde{x}$ into $X \in \mathbb{R}^{d_1\times\cdots\times d_K}$.
```

Using this notation and substituting in (17) from the main text, the objective (14) becomes

$$\Omega_{t+1} \in \arg\min_{\Omega\in\mathcal{K}_{\mathbf{p}}} \left\{ \frac{1}{2}\left\|\Omega - \left(\Omega_t - \zeta_t\left(\widetilde{S} - G_t\right)\right)\right\|_F^2 + \zeta_t \sum_{k=1}^{K} m_k\rho_k|\Psi_k|_{1,\mathrm{off}}\right\} \tag{28}$$

Expanding out the Kronecker sums, for

$$\Omega_t = \Psi_1^t \oplus \cdots \oplus \Psi_K^t, \qquad \Omega = \Psi_1 \oplus \cdots \oplus \Psi_K,$$

the Frobenius norm term in the objective (28) can be decomposed into a sum of a diagonal portion and a factor-wise sum of the off diagonal portions. This holds by Property b in Appendix A which states the off diagonal factors Ψ_k^- have disjoint support in Ω. Thus,

$$\begin{aligned}
&\left\|\Omega - \left(\Omega_t - \zeta_t\left((\widetilde{S}_1 - G_1^t) \oplus \cdots \oplus (\widetilde{S}_K - G_K^t)\right)\right)\right\|_F^2 \\
&= \left\|\left(\Psi_1 - (\Psi_1^t - \zeta_t(\widetilde{S}_1 - G_1^t))\right) \oplus \cdots \oplus \left(\Psi_K - (\Psi_K^t - \zeta_t(\widetilde{S}_K - G_K^t))\right)\right\|_F^2 \\
&= \left\|\mathrm{diag}(\Omega) - \left(\mathrm{diag}(\Omega_t) - \zeta_t \mathrm{diag}\left(\widetilde{S} - G_t\right)\right)\right\|_F^2 \\
&+ \sum_{k=1}^{K} m_k \left\|\mathrm{offd}\left(\Psi_1 - (\Psi_1^t - \zeta_t(\widetilde{S}_1 - G_1^t))\right)\right\|_F^2 .
\end{aligned}$$

Algorithm 4 TG-ISTA Implementation of TeraLasso (Detailed)

```
Input: SCM factors $S_k$, regularization parameters $\rho_i$, backtracking constant $c \in (0,1)$, initial step size $\zeta_{1,0}$, initial iterate $\Omega_{\mathrm{init}} = \Psi_1^0 \oplus \cdots \oplus \Psi_K^0$.
for $k = 1, \dots, K$ do
    $\mathbf{s}_k, U_k \leftarrow$ Eigen-decomposition of $\Psi_k^0 = U_k \mathrm{diag}(\mathbf{s}_k) U_k^T$.
    $\widetilde{S}_k \leftarrow S_k - I_{d_k} \frac{\mathrm{tr}(S_k)}{d_k} \frac{K-1}{K}$.
end for
while not converged do
    $\{\widetilde{\mathbf{s}}\}_{k=1}^K \leftarrow \mathrm{Proj}_{\mathcal{K}_\mathbf{p}}\left(\mathrm{diag}\left(\frac{1}{\mathbf{s}_1 \oplus \cdots \oplus \mathbf{s}_K}\right)\right)$.
    for $k = 1 \dots K$ do
        $G_k^t \leftarrow U_k \mathrm{diag}(\mathbf{s}_k) U_k^T$.
    end for
    for $j = 0, 1, \dots$ do
        $\zeta_t \leftarrow c^j \zeta_{t,0}$.
        for $k = 1, \dots, K$ do
            $\Psi_k^{t+1} \leftarrow \mathrm{shrink}^-_{\zeta_t \rho_k}(\Psi_k^t - \zeta_t(\widetilde{S}_k - G_k^t))$.
            Compute eigen-decomposition $U_k \mathrm{diag}(\mathbf{s}_k) U_k^T = \Psi_k^{t+1}$.
        end for
        Compute $\mathcal{Q}_{\zeta_t}(\{\Psi_k^{t+1}\}, \{\Psi_k^t\})$ via (26).
        if $f(\{\Psi_k^{t+1}\}) \leq \mathcal{Q}_{\zeta_t}(\{\Psi_k^{t+1}\}, \{\Psi_k^t\})$ as in (26) and $\min_i([\mathbf{s}_1 \oplus \cdots \oplus \mathbf{s}_K]_i) > 0$ then
            Stepsize $\zeta_t$ is acceptable; break
        end if
    end for
    Compute Barzilai-Borwein stepsize $\zeta_{t+1,0}$ via (27)
end while
Return $\{\Psi_k^{t+1}\}_{k=1}^K$.
```

Substituting into the objective (28), we obtain

$$\Omega_{t+1} \in \arg\min_{\Omega \in \mathcal{K}_\mathbf{p}} \left\{ \frac{1}{2} \left\| \mathrm{diag}(\Omega) - \left(\mathrm{diag}(\Omega_t) - \zeta_t \mathrm{diag}\left(\widetilde{S} - G_t] \right) \right) \right\|_F^2 \right.$$
$$\left. + \sum_{k=1}^K m_k \left(\frac{1}{2} \left\| \mathrm{offd}\left(\Psi_k - (\Psi_k^t - \zeta_t(\widetilde{S}_k - G_k^t)) \right) \right\|_F^2 + \zeta_t \rho_k |\Psi_k|_{1,\mathrm{off}} \right) \right\}.$$

This objective is decomposable into a sum of terms each involving either the diagonal Ω^+ or one of the off diagonal factors Ψ_k^-. Thus, we can solve for each portion of Ω independently, giving

$$\operatorname{offd}(\Psi_k^{t+1}) = \arg\min_{\operatorname{offd}(\Psi_k)} \frac{1}{2}\left\|\operatorname{offd}(\Psi_k) - \operatorname{offd}(\Psi_k^t - \zeta_t(\widetilde{S}_k - G_k^t))\right\|_F^2 + \zeta_t\rho_k|\Psi_k|_{1,\text{off}} \tag{29}$$

$$\operatorname{diag}(\Omega_{t+1}) = \arg\min_{\operatorname{diag}(\Omega)} \frac{1}{2}\left\|\operatorname{diag}(\Omega) - \operatorname{diag}\left(\Omega_t - \zeta_t\left(\widetilde{S} - G_t\right)\right)\right\|_F^2.$$

Since the diagonal $\operatorname{diag}(\Omega)$ is not regularized in (29), we have

$$\operatorname{diag}(\Omega_{t+1}) = \operatorname{diag}(\Omega_t) - \zeta_t\operatorname{diag}(\widetilde{S} - G_t),$$

i.e.

$$\operatorname{diag}(\Psi_k^{t+1}) = \operatorname{diag}(\Psi_k^t) - \zeta_t\operatorname{diag}(\widetilde{S}_k - G_k^t). \tag{30}$$

This means we can equivalently obtain the solution of the problem (29) by solving

$$\Psi_k^{t+1} = \arg\min_{\Psi_k} \frac{1}{2}\left\|\Psi_k - (\Psi_k^t - \zeta_t(\widetilde{S}_k - G_k^t))\right\|_F^2 + \zeta_t\rho_k|\Psi_k|_{1,\text{off}},$$

completing the proof.

□

2.6. Proof of Joint Convexity

Our objective function is

$$Q(\{\Psi_k\}) = -\log|\Psi_1 \oplus \cdots \oplus \Psi_K| + \langle \widehat{S}, \Psi_1 \oplus \cdots \oplus \Psi_K\rangle + \sum_k \rho_k d_k |\Psi_k|_{1,\text{off}}. \tag{31}$$

We have the following theorem. This theorem proves the joint convexity of the objective function (31) and the uniqueness of the minimizer $\widehat{\Omega}$.

THEOREM 6. *The objective function* (31) *is jointly convex in* $\{\Psi_k\}_{k=1}^K$. *Furthermore, define the set* $\mathcal{A} = \{\{\Psi_k\}_{k=1}^K | Q(\{\Psi_k\}_{k=1}^K) = Q^*\}$ *where the global minimum* $Q^* = \min_{\{\Psi_k\}_{k=1}^K} Q(\{\Psi_k\}_{k=1}^K)$. *There exists a unique* $\Omega_* \in \mathcal{K}_{\mathbf{p}}^\sharp$, *defined in* (4), *that achieves the minimum of* Q *such that*

$$\Psi_1 \oplus \cdots \oplus \Psi_K = \Omega_* \quad \forall\, \{\Psi_k\}_{k=1}^K \in \mathcal{A}. \tag{32}$$

PROOF. By definition,

$$\Psi_1 \oplus \cdots \oplus \Psi_K = \Psi_1 \otimes I_{m_1} + \cdots + I_{m_K} \otimes \Psi_K \tag{33}$$

is an affine function of $\mathbf{z} = [\mathrm{vec}(\Psi_1); \dots; \mathrm{vec}(\Psi_K)]$. Thus, since $\log |A|$ is a concave function on the space of positive definite matrices (Boyd and Vandenberghe, 2009), all the terms of Q are convex since convex functions of affine functions are convex and the elementwise ℓ_1 norm is convex. Hence Q is jointly convex in $\{\Psi_k\}_{k=1}^K$ on $\mathcal{K}_{\mathbf{p}}^\sharp$. Hence, every local minima is also global. Furthermore, for positive ρ_k at least one global minimum must exist since $|\cdot|_1$ has a global minimum at zero.

We show that a nonempty set of $\{\Psi_k\}_{k=1}^K$ such that $Q(\{\Psi_k\}_{k=1}^K)$ is minimized maps to a unique $\Omega = \Psi_1 \oplus \cdots \oplus \Psi_K$. If only one point $\{\Psi_k\}_{k=1}^K$ exists that achieves the global minimum, then the statement is proved. Otherwise, suppose that two distinct points $\{\Psi_{k,1}\}_{k=1}^K$ and $\{\Psi_{k,2}\}_{k=1}^K$ achieve the global minimum Q^*. Then, for all k define

$$\Psi_{k,\alpha} = \alpha \Psi_{k,1} + (1-\alpha)\Psi_{k,2} \tag{34}$$

By convexity, $Q(\{\Psi_{k,\alpha}\}_{k=1}^K) = Q^*$ for all $\alpha \in [0,1]$, i.e. Q is constant along the specified affine line segment. This can only be true if (up to an additive constant) the first two terms of Q are equal to the negative of the second two terms along the specified segment. Since

$$-\log |A| + \langle \widehat{S}, A \rangle \tag{35}$$

is strictly convex and smooth on the positive definite cone (i.e. the second derivative along any line never vanishes) (Boyd and Vandenberghe, 2009) and the sum of the two elementwise $\ell 1$ norms along any affine combination of variables is at most piecewise linear when smooth, this cannot hold when $\Omega_\alpha = \Psi_{1,\alpha} \oplus \cdots \oplus \Psi_{K,\alpha}$ varies with α. Hence, Ω_α must be a constant Ω^* with respect to α. Thus, the minimizing Ω^* is unique and Theorem 6 is established.

□

3. Additional experiments

3.1. Convergence of nonconvex regularization algorithm

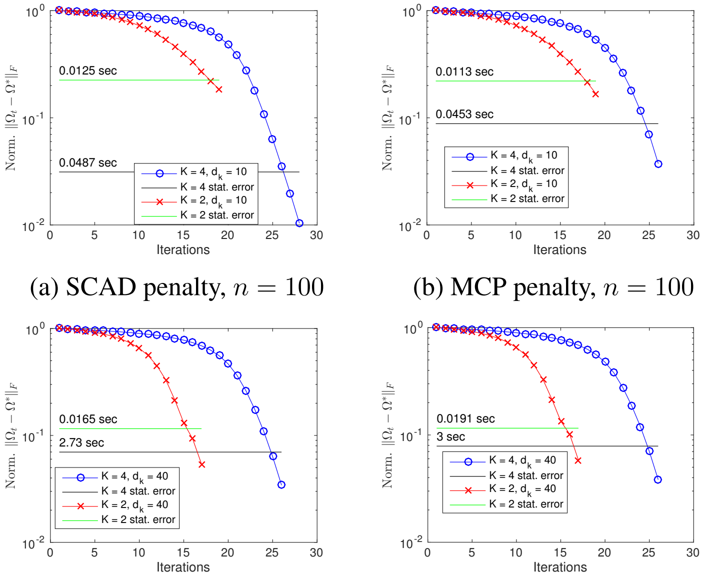
Figure 13 illustrates the convergence of the nonconvex Algorithm 2 (experiment described more thoroughly in the main text).

3.2. Computational Complexity of TG-ISTA

In Section 8, we show that TG-ISTA reaches the statistical error floor in

$$T = O_p\left(\frac{2\log K + \log(s+p) + \log\log p - \log(n \min_k m_k)}{\log\left(1 - \frac{2}{1+K^2}\right)}\right)$$

iterations.



(a) SCAD penalty, $n = 100$

(b) MCP penalty, $n = 100$

(c) SCAD penalty, $n = 1$

(d) MCP penalty, $n = 1$

Fig. 13: Geometric convergence of the nonconvex TG-ISTA implementation of TeraLasso. Shown is the normalized Frobenius norm $\|\Omega_t - \Omega^*\|_F$ of the difference between the estimate at the tth iteration and the optimal Ω^*. On the left are results comparing $K = 2$ and $K = 4$ on the same data with the same value of p (different d_k), on the right they are compared for the same value of d_k (different p). Also included are the statistical error levels, and the computation times required to reach them. Observe the consistent and rapid linear convergence rate, with logarithmic dependence on K and dimension d_k.

Each TG-ISTA iteration is also computationally efficient. Due to the representation (10), the TG-ISTA implementation of TeraLasso never needs to form the full $p \times p$ covariance. The memory footprint of the proposed implementation is $O(p + \sum_{k=1}^K d_k^2)$ as opposed to the $O(p^2)$ storage required by BiGLasso and GLasso. Since the training data itself requires $O(np)$ storage, the storage footprint of the TG-ISTA implementation of TeraLasso is scalable to large values of $p = \prod_{k=1}^K d_k$ when the d_k/p decrease in p, e.g. $d_k = p^{1/K}$. The computational cost per iteration is dominated by the computation of the gradient, which is performed by doing K eigendecompositions of size $d_1, \dots, d_K$ respectively and then computing the projection of the inverse of the Kronecker sum of the resulting eigenvalues. The former step costs $O(\sum_{k=1}^K d_k^3)$, and the second step costs $O(pK)$, giving a cost per iteration of $O\left(pK + \sum_{k=1}^K d_k^3\right)$. For $K > 1$ and $d_k/p \ll 1$, this gives a dramatic improvement on the $O(p^3) = O(\prod_{k=1}^K d_k^3)$

cost per iteration of unstructured Graphical Lasso algorithms (Guillot et al., 2012; Hsieh et al., 2014). In addition, for $K \leq 3$ the cost per iteration is comparable to the $O(d_1^3 + d_2^3 + d_3^3)$ cost per iteration of the most efficient ($K = 3$) Kronecker product GLasso methods such as Zhou (2014).

Figure 14 shows convergence speeds on various random ER graph estimation scenarios, with the BiGLasso of Kalaitzis et al. (2013) shown for comparison. Note that the BiGLasso algorithm only applies when the diagonal elements of Ω are known, so it cannot be considered to solve the general BiGLasso or TeraLasso objectives. Observe that TeraLasso's ability to efficiently exploit the Kronecker sum structure to obtain computational and memory savings allows it to quickly converge to the optimal solution, while the alternating-minimization based BiGLasso algorithm is impractically slow. All computation was timed on a 4-core, 64 bit, 2.5GHz CPU system using Matlab 2016b.

K	p	d_k	n	TeraLasso Runtime (s)	BiGLasso Runtime (s)
2	100	10	10	.0131	.84
2	625	25	10	.0147	6.81
2	2500	50	10	.0272	161
2	5625	75	10	.0401	1690
2	10^4	100	10	.0664	
2	2.5×10^5	500	10	1.62	
2	10^6	1000	10	23.2	
2	4×10^6	2000	10	427	
3	10^6	100	10	3.52	NA
3	8×10^6	200	10	11.2	NA
3	1.25×10^8	500	10	32.6	NA
3	1×10^9	1000	10	70.0	NA
4	10^4	10	10	.281	NA
4	1.6×10^5	20	10	.649	NA
4	6.25×10^6	50	10	10.8	NA
4	1.00×10^9	178	10	88.4	NA
5	1.16×10^9	65	10	124	NA

Fig. 14: Run times for the BiGLasso algorithm (Kalaitzis et al., 2013) and the proposed TG-ISTA on a $K = 2$ Kronecker sum model where the ground-truth edge topology follows a Kronecker sum Erdös-Rényi graphs for various values of the total dimension $p = d_1 d_2$ with $d_1 = d_2$. Also shown are TeraLasso results for $K = 3, 4, 5$, for which BiGLasso is not applicable. Note the 10^2 - 10^4 magnitude speed up of TeraLasso (increasing with p), allowing estimation of billion-variable covariances (10^{18} elements).

3.3. Convergence rate verification

In this section, we verify that our bounds on the rate of convergence are tight in the case of $\ell 1$ regularization. We will hold $\|\Sigma_0\|_2$ and s/p constant. We set ρ_k as in Theorem 1. By Lemma 7 in the supplement, this implies an "effective sample size" proportional to the inverse of the bound on $\|\widehat{\Omega} - \Omega_0\|_F^2/p$:

$$N_{\text{eff}} \propto (\log p)^{-1} n \min_k m_k. \tag{36}$$

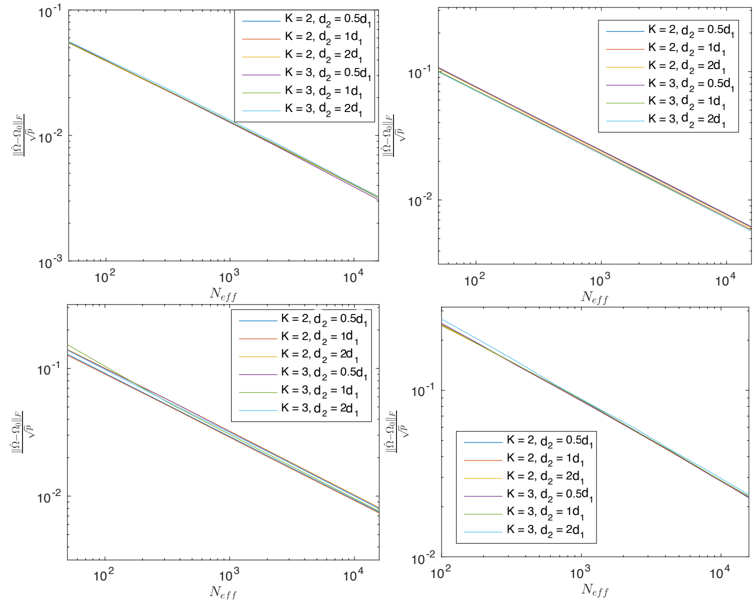


Fig. 15: Frobenius norm convergence rate for the proposed TeraLasso. Shown (ordered by increasing difficulty) are results for AR graphs with $d_1 = 40$ (top left), random ER graphs with $d_1 = 10$ (top right), $d_1 = 40$ (bottom left), and random grid graphs with $d_1 = 36$ (bottom right). For each covariance model, 6 different combinations of d_2 and K are considered, and the resulting Frobenius error is plotted versus the effective sample size N_{eff} (36).

For each experiment below, we varied K and d_2 over 6 scenarios. To ensure that the constants in the bound were minimally affected, we held Ψ_1 constant over all (K, d_2) scenarios, and let $\Psi_3 = 0$ and $d_3 = d_1$ when $K = 3$. We let d_2 vary by powers of 2, i.e. $d_2(c_d) = 2^{c_d} d_{2,\text{base}}$ where $d_{2,\text{base}}$ is a constant, allowing us to create a fixed matrix B and set $\Psi_2 = I_{d_2/d_{2,\text{base}}} \otimes B$ to ensure the eigenvalues of Ψ_2 and thus $\|\Sigma_0\|_2$ remain unaffected as d_2 (c_d) changes.

Results averaged over random training data realizations are shown in Figure 15 for ER ($d_k/2$ edges per factor), random grid ($d_k/2$ edges per factor), and AR-1 graphs (AR parameter .5 for both factors). Observe that in each case, the curves for all scenarios are very close despite the wide variation in dimensionality, indicating that our bound on the rate of convergence in Frobenius norm is tight.

3.4. Additional details for wind speed data experiments

For the wind speed data example in the main text, we first regressed out the mean for each day in the year via a 14-th order polynomial regression on the entire history from 1948-2015. As in the main text, we extracted two 20×10 spatial grids, one from eastern North America, and one from Western North America, with the latter including an expansive high-elevation area and both Atlantic and Pacific oceans (Figure 9). We compare the TeraLasso estimator to the unstructured shrinkage estimator, the non-sparse Kronecker sum estimator (TeraLasso estimator with sparsity parameter $\rho = 0$), and the Gemini sparse Kronecker product estimator of Zhou (2014). Figure 16 shows the estimated precision matrices trained on the eastern grid, using time samples from January in n years following 1948. Note the graphical structure reflects approximate auto-regressive (AR) spatial and temporal structure in each dimension. The TeraLasso estimation is much more stable than the Kronecker product estimation for small sample size n.

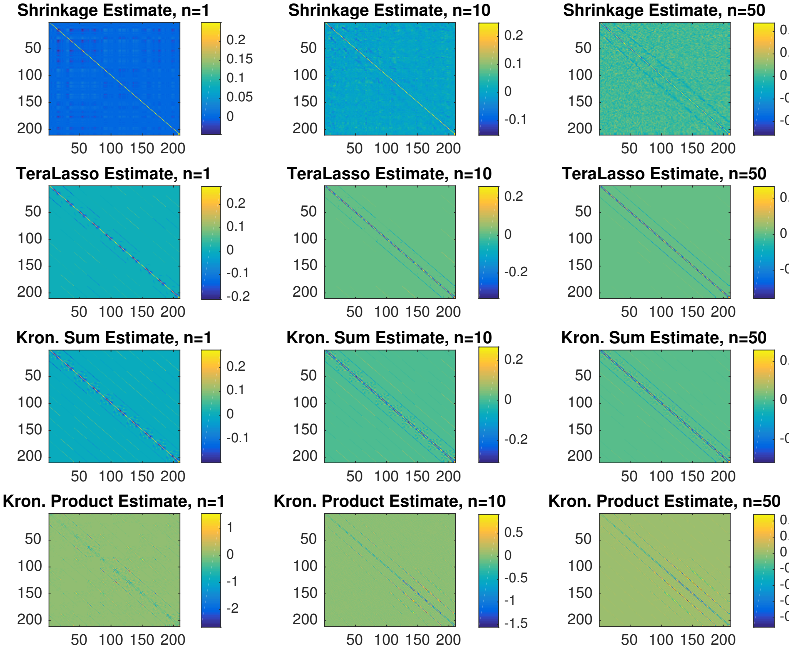


Fig. 16: Windspeed data, eastern grid. Spatial ($K = 2$) precision matrix estimation, comparing TeraLasso to unstructured and sparse Kronecker product (Gemini) techniques, using $n = 1$, 10, and 50. Observe the increasing sparsity and structure with increasing n, and TeraLasso's consistent structure even from one sample up to $n = 50$. For improved contrast, the diagonal elements have been reduced in the plot.

To quantify the fit of the estimated precision matrices to the observed wind data, we compare to an unstructured estimator in a higher sample regime. After training each estimated precision matrix (TeraLasso, Gemini, and ML Kronecker Product) on a 30-day summer interval from 1 year, as in the main experiment, we create a sample covariance $\widehat{S}_{\text{test}}$ from the same 30-day summer intervals in the remaining 50 years. We evaluate the precision matrices estimated by TeraLasso, Gemini, and ML Kronecker product using a normalized Frobenius error metric:

$$\arg\min_{\delta\in[0,1]} \|\widehat{\Omega} - (\widehat{S}_{\text{test}} + \delta I_p)^{-1}\|_F / \|(\widehat{S}_{\text{test}} + \delta I_p)^{-1}\|_F.$$

If this metric is small, the structured $\widehat{\Omega}$ is close to the unstructured $(\widehat{S}_{\text{test}} + \delta I_p)^{-1}$, indicating a good fit to the data. The small ridge δ is included to ensure that the unstructured inverse estimator $(\widehat{S}_{\text{test}} + \delta I_p)^{-1}$ is well-conditioned, with the minimum taken over δ to present the most optimistic view of Gemini and the ML Kronecker product. The results for each precision matrix are TeraLasso: 0.0728, Gemini: 0.903, and ML Kronecker Product: 0.76, confirming the superior performance of the TeraLasso estimator.

3.5. Comparison between TeraLasso and Gemini (Kronecker product) log determinant geometry

In this section, we present further analysis of the relation of the performance of TeraLasso in this wind data setting to its inherently more robust eigenstructure.

Recall the $\ell 1$ TeraLasso objective

$$-\log|\Psi_1 \oplus \cdots \oplus \Psi_K| + \langle \widehat{S}, \Psi_1 \oplus \cdots \oplus \Psi_K \rangle + \sum_{k=1}^{K} \rho_k m_k |\Psi_k|_{1,\text{off}}. \tag{37}$$

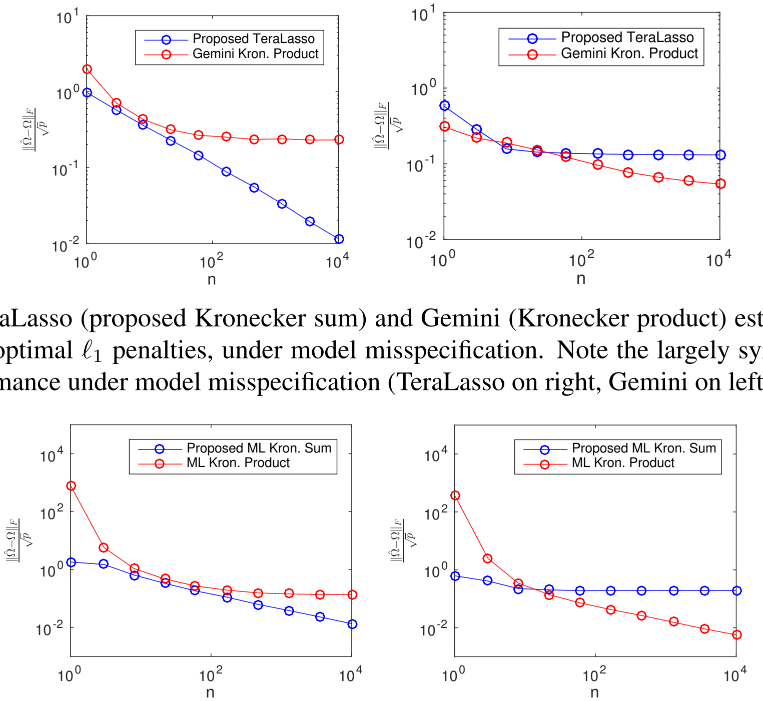
where $m_k = p/d_k$. The Gemini Kronecker product algorithm Zhou (2014) uses a similar objective function to estimate the Kronecker product covariance, which can be shown to be equivalent to

$$-\log|\Psi_1 \otimes \Psi_2| + \langle \widehat{S}, \Psi_1 \oplus \Psi_2 \rangle + \sum_{k=1}^{2} \rho_k m_k |\Psi_k|_{1,\text{off}}. \tag{38}$$

Observe that, for $K = 2$, the Gemini objective function (38) is the same as in TeraLasso objective function (37) except for the log determinant term. Figure 17 (a) compares the Kronecker product Gemini estimator to TeraLasso on data generated using precision matrix $\Psi_1 \oplus \Psi_2$, and again on data generated using the Kronecker sum precision matrix $\Psi_1 \otimes \Psi_2$, where Ψ_1, Ψ_2 are each 10×10 random ER graphs (generated as in the main text) with 5 nonzero edges. In all cases, we used the theoretically dictated optimal ℓ_1 penalty for TeraLasso from Theorem 1 in the main text and for Gemini from Theorem 3.1 in Zhou (2014).

Note that both methods perform well in the single sample regime, even under model misspecification. This apparent symmetricity is very different from the relation of the ML Kronecker sum (TeraLasso with zero penalty) and the ML Kronecker product (not directly related to Gemini), whose results on the same data are also shown in Figure 17 (b). In this case, the ML Kronecker product performs poorly in the single sample regime, whereas the ML Kronecker sum performs well in all regimes, surpassing the ML Kronecker product method in the low sample regime even when the data is generated under the Kronecker product model.

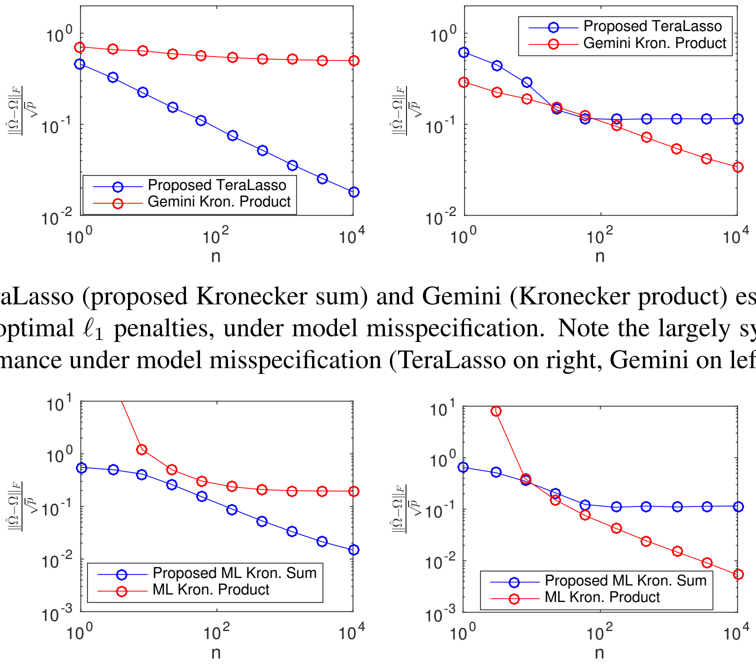
This seems to indicate that the Gemini estimator leverages some of the inherent stability of the ML Kronecker sum objective (TeraLasso) to solve the more unstable Kronecker product covariance estimation problem.



(a) TeraLasso (proposed Kronecker sum) and Gemini (Kronecker product) estimators, using optimal ℓ_1 penalties, under model misspecification. Note the largely symmetric performance under model misspecification (TeraLasso on right, Gemini on left).

(b) Gaussian maximum likelihood estimators under model misspecification. Note the significant low-sample advantage of our proposed ML Kronecker Sum estimator even under model misspecification (right).

Fig. 17: Kronecker sum and Kronecker product estimators under model misspecification. Left-hand plots were generated using Kronecker sum precision matrix $\Omega = \Psi_1 \oplus \Psi_2$, and right-hand plots were generated using Kronecker product precision matrix $\Omega = \Psi_1 \otimes \Psi_2$.



(a) TeraLasso (proposed Kronecker sum) and Gemini (Kronecker product) estimators, using optimal ℓ_1 penalties, under model misspecification. Note the largely symmetric performance under model misspecification (TeraLasso on right, Gemini on left).

(b) Gaussian maximum likelihood estimators under model misspecification. Note the significant low-sample advantage of our proposed ML Kronecker Sum estimator even under model misspecification (right).

Fig. 18: Kronecker sum and Kronecker product estimators under model misspecification, using the wind data Kronecker sum precision matrix $\Omega = \Psi_1 \oplus \Psi_2$ shown in Figure 10 (a). Left-hand plots were generated using Kronecker sum precision matrix $\Omega = \Psi_1 \oplus \Psi_2$, and right-hand plots were generated using Kronecker product precision matrix $\Omega = \Psi_1 \otimes \Psi_2$.

To further illuminate the connection between TeraLasso and Gemini, we now examine the relationship of the geometry of the differing log determinant terms. Let the eigenvalues of Ψ_k be denoted as $\lambda_{k,1}, \dots, \lambda_{k,d_k}$, and suppose that $\Psi_1 \oplus \cdots \oplus \Psi_K \succ 0$ so we can assume all the $\lambda_{k,i} \geq 0$. Using the properties of determinants and the additivity of the eigenvalues in a Kronecker sum we can write

$$\log |\Psi_1 \oplus \cdots \oplus \Psi_K| = \sum_{i_1=1}^{d_1} \cdots \sum_{i_K=1}^{d_K} \log |\lambda_{1,i_1} + \cdots + \lambda_{K,i_K}|.$$

Observe that the partial derivative of the log determinant with respect to any one eigenvalue λ_{k,i_k} is $\sum_{i_1,\dots,i_{k-1},i_{k+1},\dots i_K} 1/|\lambda_{1,i_1} + \cdots + \lambda_{K,i_K}| \leq m_k/|\lambda_{k,i_k}|$.

Correspondingly, the log determinant of a Kronecker product is

$$\log |\Psi_1 \otimes \cdots \otimes \Psi_K| = \sum_{k=1}^{K} m_k \sum_{i_k=1}^{d_k} \log |\lambda_{k,i_k}|.$$

Observe that the partial derivative of the log determinant with respect to any one eigenvalue λ_{k,i_k} is $m_k/|\lambda_{k,i_k}|$.

Thus, the geometry of the Kronecker sum log determinant term is significantly flatter than the Kronecker product log determinant, especially for larger K, indicating that the Kronecker sum estimator (TeraLasso) will enjoy more flexibility when matching the sample covariances than a Kronecker product method will.

A parallel interpretation can be obtained by recalling that the Kronecker sum of two sparse graphs is significantly sparser than the Kronecker product of the same two graphs, as discussed in the introduction of the main text.

4. Identifiable Parameterization of $\mathcal{K}_\mathbf{p}$

Observe that for any scalar 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, and we can write

$$\begin{aligned} \Psi_1 \oplus \cdots \oplus \Psi_K &= (\Psi_1 + c_1 I_{d_1}) \oplus \cdots \oplus (\Psi_K + c_K I_{d_K}) \\ &= (\Psi_1 \oplus \cdots \oplus \Psi_K) + \left(\sum_{k=1}^{K} c_k \right) I_p \\ &= \Psi_1 \oplus \cdots \oplus \Psi_K, \end{aligned} \tag{39}$$

where c_k are any scalars such that $\sum_{k=1}^{K} c_k = 0$.

The following lemma addresses this trace ambiguity, and creates an orthogonal, identifiable decomposition of Ω into factors.

Based on the original parameterization

$$B = A_1 \oplus \cdots \oplus A_K,$$

we know that the number of degrees of freedom in B is much smaller than the number of elements p^2. We thus seek a lower-dimensional parameterization of B. The Kronecker sum parameterization is not identifiable on the diagonals, so we seek a representation of B that is identifiable. In the main text, we noted that $\mathrm{diag}(B) + \mathrm{offd}(A_1) \oplus \cdots \oplus \mathrm{offd}(A_K)$ is identifiable (where $\mathrm{offd}(\mathrm{A}) = A - \mathrm{diag}(A)$), but $\mathrm{diag}(B)$ cannot be a parameter of the model since not all diagonal vectors can be expressed as a Kronecker sum. Hence while this diagonal-based decomposition is useful for stating identifiable factorwise error bounds, it is does not truly serve as a parameterization. We show in Lemma 7 that the space $\mathcal{K}_{\mathbf{p}}$ is linearly, identifiably, and orthogonally parameterized by the quantities $\left(\tau_B \in \mathbb{R}, \left\{\widetilde{A}_k \in \{A \in \mathbb{R}^{d_k \times d_k} | \mathrm{tr}(A) \equiv 0\}\right\}_{k=1}^{K}\right)$. Specifically,

LEMMA 7. *Let $B \in \mathcal{K}_{\mathbf{p}}$ and $B = A_1 \oplus \cdots \oplus A_K \in \mathcal{K}_{\mathbf{p}}$. Then B can be identifiably written as*

$$B = \tau_B I_p + (\widetilde{A}_1 \oplus \cdots \oplus \widetilde{A}_K) \tag{40}$$

where $\mathrm{tr}(\widetilde{A}_k) \equiv 0$ and the identifiable parameters $(\tau_B, \{\widetilde{A}_k\}_{k=1}^{K})$ can be computed as

$$\tau_B \;\; = \tfrac{\mathrm{tr}(B)}{p}, \qquad \widetilde{A}_k = A_k - \tfrac{\mathrm{tr}(A_k)}{d_k} I_{d_k}. \tag{41}$$

By orthogonality, the Frobenius norm can be decomposed as

$$\|B\|_F^2 = p\tau_B^2 + \sum_{k=1}^{K} m_k \|\widetilde{A}_k\|_F^2 \geq \sum_{k=1}^{K} m_k \left\| \frac{\tau_B}{K} I_{d_k} + \widetilde{A}_k \right\|_F^2,$$

noting that

$$B = \left(\frac{\tau_B}{K} I_{d_1} + \widetilde{A}_1\right) \oplus \cdots \oplus \left(\frac{\tau_B}{K} I_{d_K} + \widetilde{A}_K\right).$$

PROOF. **Part I: Identifiable Parameterization.** Let $B \in \mathcal{K}_{\mathbf{p}}$. By definition, there exists $A_1, \ldots, A_K$ such that

$$\begin{aligned} B = A_1 \oplus \cdots \oplus A_K &= \sum_{k=1}^{K} I_{[d_{1:k-1}]} \otimes A_k \otimes I_{[d_{k+1:K}]} \\ &= \sum_{k=1}^{K} \left(I_{[d_{1:k-1}]} \otimes (A_k - \tau_k I_{d_k}) \otimes I_{[d_{k+1:K}]} + \tau_k I_p\right) \\ &= \left(\sum_{k=1}^{K} \tau_k\right) I_p + ((A_1 - \tau_1 I_{d_1}) \oplus \cdots \oplus (A_K - \tau_K I_{d_K})). \end{aligned}$$

where $\tau_k = \mathrm{tr}(A_k)/d_k$. Observe that $\mathrm{tr}(A_k - \tau_k I_{d_k}) = 0$ by construction, so we can set $\widetilde{A}_k = A_k - \tau_k I_{d_k}$, creating

$$B = \left(\sum_{k=1}^{K} \tau_k\right) I_p + (\widetilde{A}_1 \oplus \cdots \oplus \widetilde{A}_K).$$

Note that in this representation, $\mathrm{tr}(\widetilde{A}_1 \oplus \cdots \oplus \widetilde{A}_K) = 0$, so letting $\tau_B = \mathrm{tr}(B)/p$,

$$\tau_B = \sum_{k=1}^{K} \tau_k,$$

and (40) in the Lemma results. It is easy to verify any B expressible in the form (40) is in $\mathcal{K}_{\mathbf{p}}$.

Thus, $(\tau_B, \{\widetilde{A}_k\}_{k=1}^K)$ parameterizes $\mathcal{K}_{\mathbf{p}}$. It remains to show that this parameterization is identifiable.

Part II: Orthogonal Parameterization. We will show that under the linear parameterization of $\mathcal{K}_{\mathbf{p}}$ by $(\tau_B, \{\widetilde{A}_k\}_{k=1}^K)$, each of the $K+1$ components are linearly independent of the others.

To see this, we compute the inner products between the components:

$$\begin{aligned}
&\langle \tau_B I_p, I_{[d_{1:k-1}]} \otimes \widetilde{A}_k \otimes I_{[d_{k+1:K}]} \rangle = \tau_B m_k \mathrm{tr}(\widetilde{A}_k) \equiv 0 \\
&\langle I_{[d_{1:k-1}]} \otimes \widetilde{A}_k \otimes I_{[d_{k+1:K}]}, I_{[d_{1:\ell-1}]} \otimes \widetilde{A}_\ell \otimes I_{[d_{\ell+1:K}]} \rangle \\
&\qquad = \mathrm{tr}\left(I_{[d_{1:k-1}]} \otimes \widetilde{A}_k \otimes I_{[d_{k+1:\ell-1}]} \otimes \widetilde{A}_\ell \otimes I_{[d_{\ell+1:K}]} \right) \\
&\qquad = \frac{p}{d_k d_\ell} \mathrm{tr}(\widetilde{A}_k) \mathrm{tr}(\widetilde{A}_\ell) \equiv 0,
\end{aligned}$$

for all $k \neq \ell$. We have recalled that by definition, $\mathrm{tr}(\widetilde{A}_k) \equiv 0$ for all k. Since all the inner products are identically zero, the components are orthogonal, thus they are linearly independent. Hence, by the definition of linear independence, this linear parameterization $(\tau_B, \{\widetilde{A}_k\}_{k=1}^K)$ is uniquely determined by $B \in \mathcal{K}_{\mathbf{p}}$ (i.e. it is identifiable).

Part III: Decomposition of Frobenius norm. Using the identifiability and orthogonality of this parameterization, we can find a direct factorwise decomposition of the Frobenius norm on $\mathcal{K}_{\mathbf{p}}$.

By orthogonality (cross term inner products equal to zero)

$$\begin{aligned}
\|B\|_F^2 &= \|\tau_B I_p\|_F^2 + \sum_{k=1}^{K} \|I_{[d_{1:k-1}]} \otimes \widetilde{A}_k \otimes I_{[d_{k+1:K}]}\|_F^2 \qquad (42) \\
&= p\tau_B^2 + \sum_{k=1}^{K} m_k \|\widetilde{A}_k\|_F^2.
\end{aligned}$$

This completes the first decomposition, representing the squared Frobenius norm as weighted sum of the squared Frobenius norms on each component.

For convenience, we also observe that given any $B \in \mathcal{K}_{\mathbf{p}}$ with identifiable parameterization

$$B = \tau_B I_p + (\widetilde{A}_1 \oplus \cdots \oplus \widetilde{A}_K),$$

we can absorb the scaled identity into the Kronecker sum and still bound the Frobenius norm decomposition. Specifically, observe that

$$p\tau_B^2 = pK \sum_{k=1}^{K} \left(\frac{\tau_B}{K}\right)^2 \geq p \sum_{k=1}^{K} \left(\frac{\tau_B}{K}\right)^2 .$$

Substituting this into (42),

$$\begin{aligned}
\|B\|_F^2 = p\tau_B^2 + \sum_{k=1}^{K} m_k \|\widetilde{A}_k\|_F^2 \geq p \sum_{k=1}^{K} \left(\frac{\tau_B}{K}\right)^2 + \sum_{k=1}^{K} m_k \|\widetilde{A}_k\|_F^2 \\
= \sum_{k=1}^{K} m_k \left(\left\| \frac{\tau_B}{K} I_{d_k} \right\|_F^2 + \|\widetilde{A}_k\|_F^2 \right) \\
= \sum_{k=1}^{K} m_k \left\| \frac{\tau_B}{K} I_{d_k} + \widetilde{A}_k \right\|_F^2 ,
\end{aligned}$$

where the last term follows because $\mathrm{tr}(\widetilde{A}_k) \equiv 0$ implies that $\langle I_{d_k}, \widetilde{A}_k \rangle \equiv 0$.

Observe that

$$B = \left(\frac{\tau_B}{K} I_{d_1} + \widetilde{A}_1\right) \oplus \cdots \oplus \left(\frac{\tau_B}{K} I_{d_K} + \widetilde{A}_K\right),$$

hence Lemma 7 is proved.

□

The identifiable parameterization of $\mathcal{K}_{\mathbf{p}}$ in Lemma 7 will provide a way to bound the spectral norm relative to the Frobenius norm. This is used to form the spectral norm bound in Theorem 2.

The following lemma is also used in the proof of Theorem 1 (cf. Proposition 18).

LEMMA 8 (SPECTRAL NORM BOUND). *For all $B \in \mathcal{K}_{\mathbf{p}}$,*

$$\|B\|_2 \leq \sqrt{\frac{K+1}{\min_k m_k}} \|B\|_F.$$

PROOF. Using the identifiable parameterization of B

$$B = \tau_B I_p + (\widetilde{A}_1 \oplus \cdots \oplus \widetilde{A}_K),$$

and the triangle inequality, we have

$$\begin{aligned}
\|B\|_2 &\le |\tau_B| + \sum_{k=1}^K \|\widetilde{A}_k\|_2 \le |\tau_B| + \sum_{k=1}^K \|\widetilde{A}_k\|_F \le \sqrt{K+1}\sqrt{\tau_B^2 + \sum_{k=1}^K \|\widetilde{A}_k\|_F^2} \\
&\le \sqrt{\frac{K+1}{\min_k m_k}}\sqrt{p\tau_B^2 + \sum_{k=1}^K m_k \|\widetilde{A}_k\|_F^2} \\
&\le \sqrt{\frac{K+1}{\min_k m_k}}\|B\|_F.
\end{aligned}$$

□

4.1. Inner Product in $\mathcal{K}_{\mathbf{p}}$

LEMMA 9 (KRONECKER SUM INNER PRODUCTS). *Suppose $B \in \mathbb{R}^{p\times p}$. Then for any $A_k \in \mathbb{R}^{d_k \times d_k}$, $k = 1, \ldots, K$,*

$$\langle B, A_1 \oplus \cdots \oplus A_K\rangle = \sum_{k=1}^K m_k \langle B_k, A_k\rangle.$$

PROOF.

$$\begin{aligned}
\langle B, A_1 \oplus \cdots \oplus A_K\rangle &= \sum_{k=1}^K \langle B, I_{[d_{1:k-1}]} \otimes A_k \otimes I_{[d_{k+1:K}]}\rangle \\
&= \sum_{k=1}^K \sum_{i=1}^{m_k} \langle B(i,i|k), A_k\rangle \\
&= \sum_{k=1}^K \left\langle \sum_{i=1}^{m_k} B(i,i|k), A_k \right\rangle \\
&= \sum_{k=1}^K m_k \langle B_k, A_k\rangle.
\end{aligned}$$

where we have used the definition of the submatrix notation $B(i,i|k)$ and the matrices $B_k = \frac{1}{m_k}\sum_{i=1}^{m_k} B(i,i|k)$. See Appendix A for the notation being used here. □

5. Proof of Theorems 1 and 2 ($\ell 1$ regularized case)

Let Ω_0 be the true value of the precision matrix Ω. Since $\Omega, \Omega_0 \in \mathcal{K}_{\mathbf{p}}$ and $\mathcal{K}_{\mathbf{p}}$ is convex, $\Delta_\Omega = \Omega - \Omega_0 \in \mathcal{K}_{\mathbf{p}}$ and we can decompose Δ_Ω into diagonal and

Kronecker sum off diagonal components:

$$\Delta_\Omega = \Omega - \Omega_0 = \mathrm{diag}(\Delta_\Omega) + (\mathrm{offd}(\Delta_{\Psi,1}) \oplus \cdots \oplus \mathrm{offd}(\Delta_{\Psi,K})), \tag{43}$$

where $\mathrm{diag}(\Delta_\Omega) = \mathrm{diag}(\Omega - \Omega_0)$ and $\mathrm{offd}(\Delta_{\Psi,k}) = \mathrm{offd}(\Psi_k - \Psi_{0,1})$. Recall that the $\mathrm{diag}(\Delta_\Omega)$ and $\mathrm{offd}(\Delta_{\Psi,k})$ terms are all identifiable given $\Delta_\Omega \in \mathcal{K}_{\mathbf{p}}$. Similarly, we can write

$$\begin{aligned} \Omega &= \mathrm{diag}(\Omega) + (\mathrm{offd}(\Psi_1) \oplus \cdots \oplus \mathrm{offd}(\Psi_K)) \\ \Omega_0 &= \mathrm{diag}(\Omega_0) + (\mathrm{offd}(\Psi_{0,1}) \oplus \cdots \oplus \mathrm{offd}(\Psi_{0,K})). \end{aligned}$$

Let $I(\cdot)$ be the indicator function. For an index set $\mathcal{A}$ and a matrix $M = [m_{ij}]$, define the operator $\mathcal{P}_{\mathcal{A}}(M) \equiv [m_{ij} I((i,j) \in \mathcal{A})]$ that projects M onto the set $\mathcal{A}$. Let $\Delta_{k,S} = \mathcal{P}_{\mathcal{S}_k}(\mathrm{offd}(\Delta_{\Psi,k}))$ be the projection of $\mathrm{offd}(\Delta_{\Psi,k})$ onto the true sparsity pattern of Ψ_k. Let $\mathcal{S}_k^c$ be the complement of $\mathcal{S}_k$, and $\Delta_{k,S^c} = \mathcal{P}_{\mathcal{S}_k^c}(\mathrm{offd}(\Delta_{\Psi,k}))$. Furthermore, let

$$\begin{aligned} \Delta_S &= (\Delta_{1,S} \oplus \cdots \oplus \Delta_{K,S}) \text{ and} \\ \Delta_{S^c} &= \Delta_{1,S^c} \oplus \cdots \oplus \Delta_{K,S^c} \end{aligned}$$

be the projection of Δ_Ω onto the sparsity set $\mathcal{S}$ and its complement. Recall neither $\mathcal{S}$ nor $\mathcal{S}^c$ includes the diagonal.

We now provide a deterministic bound on the difference in the penalty terms.

LEMMA 10. *Denote by*

$$\Delta_g := \sum_k \rho_k m_k (|\Psi_{k,0} + \Delta_{\Psi,k}|_{1,\mathrm{off}} - |\Psi_{k,0}|_{1,\mathrm{off}}),$$

Then

$$\Delta_g \geq \sum_k \rho_k m_k (|\Delta_{k,S^c}|_1 - |\Delta_{k,S}|_1) \tag{44}$$

Proof of Lemma 10. By the decomposability of the ℓ_1 norm and the reverse triangle inequality $|A + B|_1 \geq |A|_1 - |B|_1$, we have

$$\begin{aligned} &|\Psi_{k,0} + \Delta_{\Psi,k}|_{1,\mathrm{off}} - |\Psi_{k,0}|_{1,\mathrm{off}} \\ &\qquad = |\Psi_{k,0} + \Delta_{k,S}|_{1,\mathrm{off}} + |\Delta_{k,S^c}|_1 - |\Psi_{k,0}|_{1,\mathrm{off}} \\ &\qquad \geq |\Psi_{k,0}|_{1,\mathrm{off}} - |\Delta_{k,S}|_1 + |\Delta_{k,S^c}|_1 - |\Psi_{k,0}|_{1,\mathrm{off}} \\ &\qquad \geq |\Delta_{k,S^c}|_1 - |\Delta_{k,S}|_1 \end{aligned} \tag{45}$$

since $\Psi_{k,0}$ is assumed to follow sparsity pattern $\mathcal{S}_k$ by (A1). □

Let $\mathcal{A}_0$ be the event that for some constant C_0,

$$\frac{|\mathrm{tr}(\widehat{S}) - \mathrm{tr}(\Sigma_0)|}{p} \leq C_0 \|\Sigma_0\|_2 \sqrt{\frac{\log p}{pn}}; \tag{46}$$

and for each $k = 1, \dots, K$, denote by $\mathcal{A}_k$ the event such that

$$\max_{ij} \left| [S_k - \Sigma_0^{(k)}]_{ij} \right| \le C_0 \|\Sigma_0\|_2 \sqrt{\frac{\log p}{m_k n}} \tag{47}$$

holds for some absolute constant C_0 which is chosen such that probability statement in Lemma 11 holds:

LEMMA 11. *Let $\mathcal{A} = \cap_{k=0}^K \mathcal{A}_k$ as in* (47), (46). *Then* $\mathbb{P}(\mathcal{A}) \ge 1 - 2(K + 1)\exp(-c \log p)$.

Lemma 11 is proved in Section 6. Using the definition of event $\mathcal{A}$, in Section 5.2 we prove the following lemma.

LEMMA 12. *Denote by $\delta_{n,k} = C_1 \|\Sigma_0\|_2 \sqrt{\frac{\log p}{n m_k}}$. Then on event $\mathcal{A}$ the following holds: for all Δ_Ω as in* (43)

$$\left| \langle \, \mathrm{offd}(\Delta_\Omega), \widehat{S} - \Sigma_0 \, \rangle \right| \quad \le \quad \sum_{k=1}^K m_k \, |\Delta_{\Psi,k}|_{1,\mathrm{off}} \, \delta_{n,k} \tag{48}$$

where C_0 are some absolute constants.

We then have the following lemma, which we prove in Section 5.3.

LEMMA 13. *On event $\mathcal{A}$, we have for $\Delta_\Omega \in \mathcal{K}_{\mathbf{p}}$,*

$$\begin{aligned}
\left| \langle \, \mathrm{diag}(\Delta_\Omega), \widehat{S} - \Sigma_0 \, \rangle \right| \quad &\le \quad C_1 \, \|\Sigma_0\|_2 \sqrt{\frac{\log p}{n \min_k m_k}} \sqrt{(K+1)p} \, \|\mathrm{diag}(\Delta_\Omega)\|_F \\
&\le \quad \max_k \delta_{n,k} \sqrt{(K+1)p} \, \|\mathrm{diag}(\Delta_\Omega)\|_F \\
&\asymp \quad \|\Sigma_0\|_2 \, \|\mathrm{diag}(\Delta_\Omega)\|_F \max_k \sqrt{d_k} \sqrt{\frac{\log p}{n}}
\end{aligned}$$

where C_1 is an absolute constant.

5.1. Proof of Theorem 1

Let

$$G(\Delta_\Omega) = Q(\Omega_0 + \Delta_\Omega) - Q(\Omega_0) \tag{49}$$

be the difference between the objective function (10) 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 $\mathcal{K}_{\mathbf{p}}^\sharp$ (cf. Theorem 6). Define

$$\mathcal{T}_n = \left\{ \Delta_\Omega \in \mathcal{K}_{\mathbf{p}} : \Delta_\Omega = \Omega - \Omega_0, \Omega, \Omega_0 \in \mathcal{K}_{\mathbf{p}}^\sharp, \|\Delta_\Omega\|_F = M r_{n,\mathbf{p}} \right\} \tag{50}$$

where for some large enough absolute constant C to be specified,

$$r_{n,\mathbf{p}} = \frac{C\left\|\Sigma_0\right\|_2}{M}\sqrt{(s+p)\,(K+1)}\sqrt{\frac{\log p}{n\min_k m_k}} \quad \text{where} \qquad (51)$$
$$M = \frac{1}{2}\phi^2_{\max}(\Omega_0) = \frac{1}{2\phi^2_{\min}(\Sigma_0)};$$

In particular, we set $C > 9(\max_k \frac{1}{\varepsilon_k} \vee C_1)$ for C_1 as in Lemma 13.

Proposition 14 follows from Zhou et al. (2010).

PROPOSITION 14. *If $G(\Delta) > 0$ for all $\Delta \in \mathcal{T}_n$ as defined in* (50). *then $G(\Delta) > 0$ for all Δ in*

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

for $r_{n,\mathbf{p}}$ (51). *Hence if $G(\Delta) > 0$ for all $\Delta \in \mathcal{T}_n$, then $G(\Delta) > 0$ for all $\Delta \in \mathcal{T}_n \cup \mathcal{V}_n$.*

PROOF. By contradiction, suppose $G(\Delta') \leq 0$ for some $\Delta' \in \mathcal{V}_n$. Let $\Delta_0 = \frac{Mr_{n,\mathbf{p}}}{\|\Delta'\|_F}\Delta'$. Then $\Delta_0 = \theta\mathbf{0} + (1-\theta)\Delta'$, where $0 < 1-\theta = \frac{Mr_{n,\mathbf{p}}}{\|\Delta'\|_F} < 1$ by definition of Δ_0. Hence $\Delta_0 \in \mathcal{T}_n$ since by the convexity of the positive definite cone $\Omega_0 + \Delta_0 \succ 0$ because $\Omega_0 \succ 0$ and $\Omega_0 + \Delta' \succ 0$. By the convexity of $G(\Delta)$, we have that $G(\Delta_0) \leq \theta G(\mathbf{0}) + (1-\theta)G(\Delta') \leq 0$, contradicting our assumption that $G(\Delta_0) > 0$ for $\Delta_0 \in \mathcal{T}_n$. □

PROPOSITION 15. *Suppose $G(\Delta_\Omega) > 0$ for all $\Delta_\Omega \in \mathcal{T}_n$. We then have that*

$$\|\widehat{\Delta}_\Omega\|_F < Mr_{n,\mathbf{p}}.$$

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 14 (section 4.1), $\widehat{\Delta}_\Omega \notin \mathcal{T}_n \cup \mathcal{V}_n$ where $\mathcal{V}_n$ is defined therein. The proposition results. □

LEMMA 16. *Under (A1) - (A3), for all $\Delta \in \mathcal{T}_n$ for which $r_{n,\mathbf{p}} = o\left(\sqrt{\frac{\min_k m_k}{K+1}}\right)$,*

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

The proof is in Section 5.4.

By Proposition 15, it remains to show that $G(\Delta_\Omega) > 0$ on $\mathcal{T}_n$ under event $\mathcal{A}$. We show this indeed holds.

LEMMA 17. *On event $\mathcal{A}$, we have $G(\Delta) > 0$ for all $\Delta \in \mathcal{T}_n$.*

PROOF. Throughout this proof, we assume that event $\mathcal{A}$ holds. By Lemma 16, if $r_{n,\mathbf{p}} \leq \sqrt{\min_k m_k/(K+1)}$, we can write (49) using the objective (10),

$$\begin{aligned} G(\Delta_\Omega) &= \langle \Omega_0 + \Delta_\Omega, \widehat{S}\rangle - \log|\Omega_0 + \Delta_\Omega| - \langle \Omega_0, \widehat{S}\rangle + \log|\Omega_0| \qquad (52)\\ &\quad + \sum_k \rho_k m_k | \left|\Psi_{k,0} + \Delta_{\Psi,k}\right|_{1,\text{off}} - \sum_k \rho_k m_k \left|\Psi_{k,0}\right|_{1,\text{off}} \\ &\geq \langle \Delta_\Omega, \widehat{S}\rangle - \langle \Delta_\Omega, \Sigma_0\rangle + \frac{2}{9\|\Omega_0\|_2^2}\|\Delta_\Omega\|_F^2 \\ &\quad + \sum_k \rho_k m_k (|\Psi_{k,0} + \Delta_{\Psi,k}|_{1,\text{off}} - |\Psi_{k,0}|_{1,\text{off}}) \\ &= \langle\, \text{diag}(\Delta_\Omega), \widehat{S} - \Sigma_0 \,\rangle \; + \; \langle\, \text{offd}(\Delta_\Omega), \widehat{S} - \Sigma_0 \,\rangle \; + \frac{2}{9\|\Omega_0\|_2^2}\|\Delta_\Omega\|_F^2 \\ &\quad \underbrace{+ \sum_k \rho_k m_k (|\Psi_{k,0} + \Delta_{\Psi,k}|_{1,\text{off}} - |\Psi_{k,0}|_{1,\text{off}})\,.}_{\Delta_g} \end{aligned}$$

We next bound the inner product term under event $\mathcal{A}$. Substituting the bound of Lemma 12 and (44) into (52), under event $\mathcal{A}$, we have by choice of $\rho_k = \delta_{n,p}/\varepsilon_k$ where $0 < \varepsilon_k < 1$ for all k,

$$\begin{aligned} &\sum_{k=1}^K m_k \rho_k \left(\left|\Psi_k + \Delta_{\Psi,k}\right|_{1,\text{off}} - \left|\Psi_k\right|_{1,\text{off}} \right) + \; \langle\, \text{offd}(\Delta_\Omega), \widehat{S} - \Sigma_0 \,\rangle \\ &\geq \sum_{k=1}^K m_k \rho_k \left(\left|\Delta_{k,S^c}\right|_1 - \left|\Delta_{k,S}\right|_1 \right) - \sum_{k=1}^K m_k \left|\Delta_{\Psi,k}\right|_{1,\text{off}} \delta_{n,k} \\ &\geq \sum_{k=1}^K m_k \rho_k \left(\left|\Delta_{k,S^c}\right|_1 - \left|\Delta_{k,S}\right|_1 \right) - \sum_{k=1}^K m_k \delta_{n,k} \left(\left|\Delta_{k,S^c}\right|_1 + \left|\Delta_{k,S}\right|_1 \right) \\ &\geq -2 \max_k \rho_k \sum_{k=1}^K m_k \left|\Delta_{k,S}\right|_1 = -2 \max_k \rho_k \left|\Delta_{\Omega,S}\right|_1 \end{aligned}$$

For the diagonal part, we have by Lemma 13

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

we have for all $\Delta_\Omega \in \mathcal{T}_n$, and $C'' = \max_k(\frac{2}{\varepsilon_k}) \vee \sqrt{2}C_1$, and for $K \geq 1$,

$$\begin{aligned}
G \quad (\Delta_\Omega) &\geq \langle \operatorname{diag}(\Delta_\Omega), \widehat{S} - \Sigma_0 \rangle \; - 2 \max_k \rho_{n,k} \left|\Delta_{\Omega,S}\right|_1 + \frac{2}{9\|\Omega_0\|_2^2}\|\Delta_\Omega\|_F^2 \\
&> \frac{2}{9\|\Omega_0\|_2^2}\|\Delta_\Omega\|_F^2 \\
&- \max_k \delta_{n,k} \left(\sqrt{p}\sqrt{K+1} \left\|\operatorname{diag}(\Delta_\Omega)\right\|_F + 2 \max_k \frac{1}{\varepsilon_k} \left|\Delta_{\Omega,S}\right|_1 \right) \\
&\geq \frac{2}{9\|\Omega_0\|_2^2}\|\Delta_\Omega\|_F^2 - C' \left\|\Sigma_0\right\|_2 \sqrt{\frac{\log p}{n \min_k m_k}} \cdot \\
&\quad \left(\sqrt{(K+1)p} \left\|\operatorname{diag}(\Delta_\Omega)\right\|_F + \sqrt{2s}\|\Delta_{\Omega,S}\|_F \right) \\
&\geq \frac{2}{9\|\Omega_0\|_2^2} \left\|\Delta_\Omega\right\|_F^2 \\
&- C' \left(\sqrt{2}\sqrt{(K+1)(p+s)} \left\|\Delta_\Omega\right\|_F \right) \left\|\Sigma_0\right\|_2 \sqrt{\frac{\log p}{n \min_k m_k}} \\
&= \|\Delta\|_F^2 \left(\frac{2}{9\left\|\Omega_0\right\|_2^2} - C'' \left\|\Sigma_0\right\|_2 \sqrt{\frac{\log p}{n \min_k m_k}} \frac{\sqrt{(K+1)(p+s)}}{M r_{n,\mathbf{p}}} \right) > 0
\end{aligned}$$

which holds for all $\Delta_\Omega \in \mathcal{T}_n$, where we use the following bounds: for all $K \geq 1$.

$$\sqrt{(K+1)p} \left\|\operatorname{diag}(\Delta_\Omega)\right\|_F + \sqrt{2s}\|\Delta_{\Omega,S}\|_F \quad \leq \quad \sqrt{2}\sqrt{(K+1)(p+s)} \left\|\Delta_\Omega\right\|_F$$

and

$$\begin{aligned}
&C'' \left\|\Sigma_0\right\|_2 \sqrt{\frac{\log p}{n \min_k m_k}} \sqrt{(K+1)(p+s)} \frac{1}{M r_{n,\mathbf{p}}} \\
&= \frac{C''}{CM} = \frac{2C''}{C}\phi_{\min}^2(\Sigma_0) < \frac{2}{9\left\|\Omega_0\right\|_2^2}
\end{aligned}$$

where $M = \frac{1}{2\phi^2_{\min}(\Sigma_0)}$, which holds so long as C is chosen to be large enough in

$$r_{n,\mathbf{p}} = C \left\|\Sigma_0\right\|_2 \sqrt{(s+p)(K+1)\frac{\log p}{n \min_k m_k}};$$

For example, we set $C > 9C'' = 9(\max_k(\frac{2}{\varepsilon_k}) \vee \sqrt{2}C_1)$.

Theorem 1 follows from Proposition 15 immediately. □

5.2. Proof of Lemma 12

Assume that the event $\mathcal{A}$ of Lemma 11 holds. Using the definition of Δ_Ω (43), the projection operator $\operatorname{Proj}_{\widetilde{K}_\mathbf{p}}(\cdot)$, and letting $\tau_\Sigma = (K-1)\frac{\operatorname{tr}(\widehat{S}-\Sigma_0)}{p}$, we have

$$\left|\langle \mathrm{offd}(\Delta_\Omega), \widehat{S} - \Sigma_0\rangle\right| = |\langle \Delta_\Omega, \mathrm{Proj}_{\widetilde{K}_{\mathbf{p}}}(\widehat{S} - \Sigma_0)\rangle| \tag{53}$$
$$= \left|\langle \mathrm{offd}(\Delta_\Omega), (S_1 - \Sigma_0^{(1)}) \oplus \cdots \oplus (S_K - \Sigma_0^{(K)}) - \tau_\Sigma I_p\rangle\right|$$
$$= \left|\langle \mathrm{offd}(\Delta_{\Psi,1}) \oplus \cdots \oplus \mathrm{offd}(\Delta_{\Psi,K}), (S_1 - \Sigma_0^{(1)}) \oplus \cdots \oplus (S_K - \Sigma_0^{(K)})\rangle\right|,$$

where we have used the fact that $\mathrm{offd}(\Delta_{\Psi,1}) \oplus \cdots \oplus \mathrm{offd}(\Delta_{\Psi,K})$ is zero along the diagonal and thus has zero inner product with I_p. Substituting Lemma 13 and the definitions of subevents under $\mathcal{A}$, we have by (54) and Lemma 9,

$$\begin{aligned}
\left|\langle \mathrm{offd}(\Delta_\Omega), \widehat{S} - \Sigma_0\rangle\right| &= \sum_{k=1}^K m_k |\langle \mathrm{offd}(\Delta_{\Psi,k}), S_k - \Sigma_0^{(k)}\rangle| \qquad (54)\\
&\leq \sum_{k=1}^K m_k \sum_{i,j=1}^{d_k} |[\mathrm{offd}(\Delta_{\Psi,k})]_{ij}| \cdot \max_{ij} \left|[S_k - \Sigma_0^{(k)}]_{ij}\right|\\
&\leq C\|\Sigma_0\|_2 \sum_{k=1}^K m_k |\Delta_{\Psi,k}|_{1,\mathrm{off}} \sqrt{\frac{\log p}{m_k n}}.
\end{aligned}$$

□

5.3. *Proof of Lemma 13: Bound on Inner Product for Diagonal*

Let $\widetilde{\Delta}_\Omega = \Delta_\Omega - \tau_\Omega I_p$. Recall the identifiable parameterization of Δ_Ω (Lemma 7)

$$\Delta_\Omega = \tau_\Omega I_p + \widetilde{\Delta}_{\Psi,1} \oplus \cdots \oplus \widetilde{\Delta}_{\Psi,K}$$

where $\tau_\Omega = \mathrm{tr}(\Delta_\Omega)/p$ and $\widetilde{\Delta}_{\Psi,k}$ are given in the lemma. We then have $\mathrm{tr}(\widetilde{\Delta}_{\Psi,j}) = 0$ and

$$\sum_{k=1}^K \left\|\mathrm{diag}(\widetilde{\Delta}_{\Psi,k})\right\|_F^2 m_k + p\tau_\Omega^2 = \|\mathrm{diag}(\Delta_\Omega)\|_F^2 \tag{55}$$

by othogonality of the decomposition. By Lemma 9, we can write

$$\begin{aligned}
\left| \langle\, \mathrm{diag}(\widetilde{\Delta}_\Omega), \widehat{S} - \Sigma_0 \,\rangle \right| &\leq \sum_{k=1}^K m_k |\, \langle\, S_k - \Sigma_0^{(k)}, \mathrm{diag}(\widetilde{\Delta}_{\Psi_k}) \,\rangle\, |\\
&\leq C\,\|\Sigma_0\|_2 \sum_{k=1}^K m_k \left|\mathrm{diag}(\widetilde{\Delta}_{\Psi,k})\right|_1 \sqrt{\frac{\log p}{n m_k}}\\
&\leq C\,\|\Sigma_0\|_2 \sum_{k=1}^K \sqrt{m_k}\sqrt{d_k} \left\|\mathrm{diag}(\widetilde{\Delta}_{\Psi,k})\right\|_F \sqrt{\frac{\log p}{n}}.
\end{aligned}$$

Moreover, under $\mathcal{A}_0$, we have

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

Summing these terms together, we have

$$\begin{aligned}
&\left| \left\langle \operatorname{diag}(\Delta_\Omega), \widehat{S} - \Sigma_0 \right\rangle \right| \\
&\leq \;\; C_0 \|\Sigma_0\|_2 \sqrt{\frac{\log p}{n}} \left(\sum_{k=1}^{K} \sqrt{m_k}\sqrt{d_k} \left\| \operatorname{diag}(\widetilde{\Delta}_{\Psi,k}) \right\|_F + |\tau_\Omega|\sqrt{p} \right) \\
&\leq \;\; C_0 \max_k \sqrt{d_k}\, \|\Sigma_0\|_2 \sqrt{\frac{\log p}{n}} \sqrt{K+1}\, \|\operatorname{diag}(\Delta_\Omega)\|_F \qquad (56) \\
&= \;\; C_0 \max_k \left(\sqrt{\frac{\log p}{n m_k}}\, \|\Sigma_0\|_2 \right) \sqrt{(K+1)p}\, \|\operatorname{diag}(\Delta_\Omega)\|_F \\
&= \;\; C_0 \sqrt{\frac{\log p}{n \min_k m_k}}\, \|\Sigma_0\|_2 \sqrt{(K+1)p}\, \|\operatorname{diag}(\Delta_\Omega)\|_F \\
&\asymp \;\; \max_k \delta_{n,k} \sqrt{(K+1)p}\, \|\operatorname{diag}(\Delta_\Omega)\|_F
\end{aligned}$$

where in (56), we have used the following inequality in view of (55):

$$\left(\sum_{k=1}^{K} \sqrt{m_k} \left\| \operatorname{diag}(\widetilde{\Delta}_{\Psi,k}) \right\|_F + |\tau_\Omega|\sqrt{p} \right) \leq \sqrt{K+1}\, \|\operatorname{diag}(\Delta_\Omega)\|_F \, .$$

□

5.4. *Proof of Lemma 16*

We first state Proposition 18

PROPOSITION 18. *Under (A1)-(A3), for all* $\Delta \in \mathcal{T}_n$,

$$\phi_{\min}(\Omega_0) > 2M r_{n,\mathbf{p}} \sqrt{\frac{K+1}{\min_k m_k}} \geq \|\Delta\|_2 / 2 \qquad (57)$$

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. We first show that (57) holds for $\Delta \in \mathcal{T}_n$. Indeed, by Corollary 8, we have for all $\Delta \in \mathcal{T}_n$

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

so long as

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

where $\kappa(\Sigma_0) = \phi_{\max}(\Sigma_0)/\phi_{\min}(\Sigma_0)$ is the condition number of Σ_0.

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 $\varepsilon < 1$,

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

given that by definition of $\mathcal{T}_n$ and (57).

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

$$\begin{aligned} \log|\Omega_0 + \Delta| - \log|\Omega_0| = {}& \frac{d}{dv}\log|\Omega_0 + v\Delta|\Big|_{v=0} \Delta \\ &+ \int_0^1 (1-v)\frac{d^2}{dv^2}\log|\Omega_0 + v\Delta| dv \\ = {}& \langle \Sigma_0, \Delta \rangle - a. \end{aligned} \tag{58}$$

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, suppose that

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

where (57), we have for all $\Delta \in \mathcal{T}_n$,

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

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

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

Hence,

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

Thus, substituting into (58), the lemma is proved. □

5.5. Proof of Theorem 2: Factorwise and Spectral Norm Bounds

PROOF. **Part I: Factor-wise bound**. From the proof of Theorem 1, we know that under event $\mathcal{A}$,

$$\|\Delta_\Omega\|_F^2 \leq c(K+1)(s+p)\frac{\log p}{n\min_k m_k}. \tag{59}$$

Furthermore, since the identifiable parameterizations of $\widehat{\Omega}, \Omega_0$ are of the form (41) by construction in Lemma 7)

$$\widehat{\Omega} = \widehat{\tau} I_p + (\widetilde{\Psi}_1 \oplus \cdots \oplus \widetilde{\Psi}_K)$$
$$\Omega_0 = \tau_0 I_p + (\widetilde{\Psi}_{0,1} \oplus \cdots \oplus \widetilde{\Psi}_{0,K}),$$

we have that the identifiable parameterization of Δ_Ω is

$$\Delta_\Omega = \tau_\Delta I_p + (\widetilde{\Delta}_1 \oplus \cdots \oplus \widetilde{\Delta}_K), \tag{60}$$

where $\tau_\Delta = \widehat{\tau} - \tau_0$, $\widetilde{\Delta}_k = \widetilde{\Psi}_k - \widetilde{\Psi}_{0,k}$. Observe that $\mathrm{tr}(\widetilde{\Delta}_k) = \mathrm{tr}(\widetilde{\Psi}_k) - \mathrm{tr}(\widetilde{\Psi}_{0,k}) = 0$.

By Lemma 7 then,

$$\|\Delta_\Omega\|_F^2 = p\tau_\Delta^2 + \sum_{k=1}^K m_k \|\widetilde{\Delta}_k\|_F^2.$$

Thus, the estimation error on the underlying parameters is bounded by (59)

$$p\tau_\Delta^2 + \sum_{k=1}^K m_k \|\widetilde{\Delta}_k\|_F^2 \le c(K+1)(s+p)\frac{\log p}{n \min_k m_k},$$

or, dividing both sides by p

$$\begin{aligned} \tau_\Delta^2 + \sum_{k=1}^K \frac{\|\widetilde{\Delta}_k\|_F^2}{d_k} &\le c(K+1)\frac{s+p}{p}\frac{\log p}{n \min_k m_k} \\ &= c(K+1)\left(\frac{s}{p}+1\right)\frac{\log p}{n \min_k m_k}. \end{aligned} \tag{61}$$

Recall that $s = \sum_{k=1}^K m_k s_k$, so $\frac{s}{p} = \sum_{k=1}^K \frac{s_k}{d_k}$. Substituting into (61)

$$\tau_\Delta^2 + \sum_{k=1}^K \frac{\|\widetilde{\Delta}_k\|_F^2}{d_k} \le c(K+1)\left(1 + \sum_{k=1}^K \frac{s_k}{d_k}\right)\frac{\log p}{n \min_k m_k}. \tag{62}$$

From this, it can be seen that the bound converges as the m_k increase with constant K. To put the bound in the form stated in the theorem, note that since

$\tau_\Delta I_p + (\widetilde{\Delta}_1^+ \oplus \cdots \oplus \widetilde{\Delta}_K^+)$

$$\begin{aligned}
\frac{\|\mathrm{diag}(\Delta_\Omega)\|_2^2}{\max_k d_k} &\le \frac{\left(\tau_\Delta + \sum_{k=1}^K \|\widetilde{\Delta}_k^+\|_2\right)^2}{\max_k d_k} \\
&\le \frac{K+1}{\max_k d_k}\left(\tau_\Delta^2 + \sum_{k=1}^K \|\mathrm{diag}(\widetilde{\Delta}_k)\|_2^2\right) \\
&\le (K+1)\left(\tau_\Delta^2 + \sum_{k=1}^K \frac{\|\mathrm{diag}(\widetilde{\Delta}_k)\|_F^2}{d_k}\right).
\end{aligned}$$

Part II: Spectral norm bound. The factor-wise bound immediately implies the bound on the spectral norm $\|\Delta_\Omega\|_2$ of the error under event A. We recall the identifiable representation (60)

$$\Delta_\Omega = \tau_\Delta I_p + (\widetilde{\Delta}_1 \oplus \cdots \oplus \widetilde{\Delta}_K).$$

By Property ci in Appendix A and the fact that the spectral norm is upper bounded by the Frobenius norm,

$$\begin{aligned}
\|\Delta_\Omega\|_2 &\le |\tau_\Delta| + \sum_{k=1}^K \|\widetilde{\Delta}_k\|_2 \le |\tau_\Delta| + \sum_{k=1}^K \|\widetilde{\Delta}_k\|_F \\
&\le \sqrt{K+1}\sqrt{\tau_\Delta^2 + \sum_{k=1}^K \|\widetilde{\Delta}_k\|_F^2} \\
&\le \sqrt{K+1}\sqrt{\max_k d_k}\sqrt{\tau_\Delta^2 + \sum_{k=1}^K \frac{\|\widetilde{\Delta}_k\|_F^2}{d_k}} \\
&\le c(K+1)\sqrt{(\max_k d_k)\left(1 + \sum_{k=1}^K \frac{s_k}{d_k}\right)}\sqrt{\frac{\log p}{n \min_k m_k}},
\end{aligned}$$

where in the second line, we have used the fact that for a_k elements of $\mathbf{a} \in \mathbb{R}^K$ the norm relation $\|\mathbf{a}\|_1 \le \sqrt{K}\|\mathbf{a}\|_2$ implies $(\sum_{k=1}^K |a_k|) \le \sqrt{K}\sqrt{\sum_{k=1}^K a_k^2}$. □

6. Proof of Lemma 11: Subgaussian Concentration

We first state the following concentration result, proved in Section 6.1. Recall that $m_k = p/d_k$.

LEMMA 19 (SUBGAUSSIAN CONCENTRATION). *Suppose that* $\log p \ll m_k n$ *for all* k. *Then, with probability at least* $1 - 2\exp(-c' \log p)$,

$$|\langle \Delta, S_k - \Sigma_0^{(k)} \rangle| \leq C |\Delta|_1 \|\Sigma_0\|_2 \sqrt{\frac{\log p}{m_k n}}$$

for all $\Delta \in \mathbb{R}^{d_k \times d_k}$, *where* c' *is a constant depending on* C *given in the proof.*

We can now prove Lemma 11.

PROOF. By Lemma 19 we have that event $\mathcal{A}_k$ (47), i.e. the event that

$$\max_{ij} \left| [S_k - \Sigma_0^{(k)}]_{ij} \right| = \max_{ij} \left| \langle \mathbf{e}_i \mathbf{e}_j^T, S_k - \Sigma_0^{(k)} \rangle \right| \leq C \|\Sigma_0\|_2 \sqrt{\frac{\log p}{m_k n}},$$

holds with probability at least $1 - 2\exp(-c' \log p)$.

Note that $\mathbb{E}[\mathrm{tr}(\widehat{S})] = \mathrm{tr}(\Sigma_0)$. Viewing $\frac{1}{p}\mathrm{tr}(\Sigma_0)$ as a 1×1 covariance factor since $\frac{1}{p}\mathrm{tr}(\widehat{S}) = \frac{1}{pn} \sum_{i=1}^n \mathrm{vec}(X_i)\mathrm{vec}(X_i)^T$, we can invoke the proof of Lemma 19 and show that with probability at least $1 - 2\exp(-c' \log p)$ the event $\mathcal{A}_0$ (46) will hold. Recall that $\mathcal{A} = \mathcal{A}_0 \cap \mathcal{A}_1 \cap \cdots \cap \mathcal{A}_K$. By the union bound, we have $\mathbb{P}(\mathcal{A}) \geq 1 - 2(K+1)\exp(-c \log p)$. □

6.1. Proof of Lemma 19

Define a K-way generalization of the invertible Pitsianis-Van Loan type (Van Loan and Pitsianis, 1993) rearrangement operator $\mathcal{R}_k(\cdot)$, which maps $p \times p$ matrices to $d_k^2 \times m_k^2$ matrices. For a matrix $M \in \mathbb{R}^{p \times p}$ we set

$$\begin{aligned} \mathcal{R}_k(M) &= [\ \mathbf{m}_1 \quad \ldots \quad \mathbf{m}_{m_k^2}\ ], \\ \mathbf{m}_{(i-1)m_k + j} &= \mathrm{vec}(M(i,j|k)), \end{aligned} \tag{63}$$

where we use the $M(i,j|k) \in \mathbb{R}^{d_k \times d_k}$ subblock notation (see Section 2 in the main text). Using this notation, we have the following concentration result.

LEMMA 20. *Let* $\mathbf{u} \in S^{d_k^2 - 1}$ *and* $\mathbf{f} = \mathrm{vec}(I_{m_k})$. *Assume that* $\mathbf{x}_t = \Sigma_0^{1/2} \mathbf{z}_t$ *where* $\mathbf{z}_t$ *has independent entries* $z_{t,f}$ *such that* $\mathbb{E} z_{t,f} = 0$, $\mathbb{E} z_{t,f}^2 = 1$, *and* $\|z_{t,f}\|_{\psi_2} \leq K$. *Let* $\Delta_n = \widehat{S} - \Sigma_0$. *Then for all* $0 \leq \frac{\epsilon}{\sqrt{m_k}} < \frac{1}{2}$:

$$\mathbb{P}(|\mathbf{u}^T \mathcal{R}_k(\Delta_n)\mathbf{f}| \geq \epsilon \sqrt{m_k} \|\Sigma_0\|_2) \leq 2\exp\left(-c \frac{\epsilon^2 n}{K^4}\right)$$

where c *is an absolute constant and* $\|\cdot\|_{\psi_2}$ *is the subgaussian norm.*

PROOF. We prove the lemma for $k = 1$. The proof for the remaining k follow similarly.

By the definition (63) of the permutation operator $\mathcal{R}_1$ and letting $\mathbf{x}_t(i) = [x_{t,(i-1)m_1+1}, \dots, x_{t,im_1}]$,

$$\mathcal{R}_1(\widehat{S}) = \frac{1}{n}\sum_{t=1}^{n} \begin{bmatrix} \mathrm{vec}(\mathbf{x}_t(1)\mathbf{x}_t(1)^T)^T \\ \mathrm{vec}(\mathbf{x}_t(1)\mathbf{x}_t(2)^T)^T \\ \vdots \\ \mathrm{vec}(\mathbf{x}_t(d_1)\mathbf{x}_t(d_1)^T)^T \end{bmatrix} \tag{64}$$

Hence,

$$\mathbf{u}^T\mathcal{R}_1(\widehat{S})\mathbf{f} = \frac{1}{n}\sum_{t=1}^{n} \mathbf{x}_t^T(U \otimes I_{m_k})\mathbf{x}_t = \frac{1}{n}\sum_{t=1}^{n} \mathbf{z}_t^T M \mathbf{z}_t \tag{65}$$

where $M = \Sigma_0^{1/2}(U \otimes I_{m_k})\Sigma_0^{1/2}$, $U = \mathrm{vec}_{d_1,d_1}^{-1}(\mathbf{u})$.

Thus, by the Hanson-Wright inequality (Rudelson et al., 2013),

$$\begin{aligned} &\mathbb{P}(|\mathbf{u}^T\mathcal{R}_1(\widehat{S})\mathbf{f} - \mathbb{E}[\mathbf{u}^T\mathcal{R}_1(\widehat{S})\mathbf{f}]| \geq \tau) \qquad (66)\\ &\leq 2\exp\left[-c\min\left(\frac{\tau^2N^2}{K^4 n\|M\|_F^2}, \frac{\tau n}{K^2\|M\|_2}\right)\right] \\ &\leq 2\exp\left[-c\min\left(\frac{\tau^2N}{K^4 m_1\|\Sigma_0\|_2^2}, \frac{\tau n}{K^2\|\Sigma_0\|_2}\right)\right] \end{aligned}$$

since $\|U \otimes I_{m_1}\|_2 = \|U\|_2 \leq 1$ and $\|U \otimes I_{m_1}\|_F^2 = \|U\|_F^2\|I_{m_1}\|_F^2 = m_1$. Substituting $\epsilon = \frac{\tau}{\sqrt{m_1}\|\Sigma_0\|_2}$

$$\mathbb{P}(|\mathbf{u}^T\mathcal{R}_1(\Delta_n)\mathbf{f}| \geq \epsilon\sqrt{m_1}\|\Sigma_0\|_2) \leq 2\exp\left(-c\frac{\epsilon^2 n}{K^4}\right) \tag{67}$$

for all $\frac{\epsilon^2 n}{K^4} \leq \frac{\epsilon n\sqrt{m_1}}{K^2}$, i.e. $\epsilon \leq K^2\sqrt{m_1} \leq \frac{\sqrt{m_1}}{2}$, since $K^2 > \frac{1}{2}$ by definition.
□

We can now prove Lemma 19.

PROOF. Consider the inner product $\langle \Delta, S_k - \Sigma_0^{(k)} \rangle$, where Δ is an arbitrary $d_k \times d_k$ matrix. Let

$$\mathbf{h} = \mathrm{vec}(\Delta), \qquad \mathbf{f} = \mathrm{vec}(I_{m_k \times m_k}).$$

By the definition of the factor covariances S_k and the rearrangement operator $\mathcal{R}_k$, it can be seen that

$$\mathrm{vec}(S_k) = \frac{1}{m_k}\mathcal{R}_k(\widehat{S})\mathbf{f},$$

and that similarly by the definition of the factor covariances $\Sigma_0^{(k)}$

$$\mathrm{vec}(\Sigma_0^{(k)}) = \frac{1}{m_k}\mathcal{R}_k(\Sigma_0)\mathbf{f}.$$

Hence,

$$\begin{aligned}\langle \Delta, S_k - \Sigma_0^{(k)} \rangle &= \frac{1}{m_k} \langle \mathrm{vec}(\Delta), \mathcal{R}_k(\widehat{S} - \Sigma_0)\mathbf{f} \rangle \qquad (68)\\ &= \frac{1}{m_k} \mathbf{h}^T \mathcal{R}_k(\widehat{S} - \Sigma_0)\mathbf{f} \\ &= \frac{1}{m_k} \sum_{i=1}^{d_k^2} h_i \mathbf{e}_i^T \mathcal{R}_k(\widehat{S} - \Sigma_0)\mathbf{f}\end{aligned}$$

by the linearity of the rearrangement operator, the definition of the inner product, and the definition of the unit vector $\mathbf{e}_i$ as the i-th column of the $d_k^2 \times d_k^2$ identity matrix.

We can apply Lemma 20 and take a union bound over $i = 1, \dots, d_k^2$. By Lemma 20,

$$\mathbb{P}\left(\left| \mathbf{e}_i^T \mathcal{R}_k(\widehat{S} - \Sigma_0)\mathbf{f} \right| \geq \epsilon \sqrt{m_k} \|\Sigma_0\|_2 \right) \leq 2 \exp\left(-c \frac{\epsilon^2 n}{K^4} \right)$$

for $0 \leq \frac{\epsilon}{\sqrt{m_k}} \leq \frac{1}{2}$. Taking the union bound over all i, we obtain

$$\begin{aligned}\mathbb{P}\left(\max_i |\mathbf{e}_i^T \mathcal{R}_k(\widehat{S} - \Sigma_0)\mathbf{f}| \geq \epsilon \|\Sigma_0\|_2 \sqrt{m_k} \right) &\leq 2 d_k^2 \exp\left(-c \frac{\epsilon^2 n}{K^4} \right) \\ &\leq 2 \exp\left(2 \log d_k - c \frac{\epsilon^2 n}{K^4} \right).\end{aligned}$$

Setting $\epsilon = C\sqrt{\frac{\log p}{n}}$ for large enough C and recalling that $m_k = p/d_k$, with probability at least $1 - 2\exp(-c' \log p)$ we have

$$\max_i |\mathbf{e}_i^T \mathcal{R}_k(\widehat{S} - \Sigma_0)\mathbf{f}| \leq C \|\Sigma_0\|_2 \sqrt{m_k} \sqrt{\frac{\log p}{n}}$$

where we assume $\log p \leq \frac{n m_k}{4C^2}$ and let $c' = \frac{cC^2}{K^4} - 2$. Hence, by (68)

$$\begin{aligned}|\langle \Delta, S_k - \Sigma_0^{(k)} \rangle| &= \frac{1}{m_k} \left| \sum_{i=1}^{d_k^2} h_i \mathbf{e}_i^T \mathcal{R}_k(\widehat{S} - \Sigma_0)\mathbf{f} \right| \\ &\leq \frac{1}{m_k} \sum_{i=1}^{d_k^2} |h_i \mathbf{e}_i^T \mathcal{R}_k(\widehat{S} - \Sigma_0)\mathbf{f}| \\ &\leq C \|\Sigma_0\|_2 \frac{1}{\sqrt{m_k}} \sqrt{\frac{\log p}{n}} \sum_{i=1}^{d_k^2} h_i \\ &= C \|\Sigma_0\|_2 \sqrt{\frac{\log p}{m_k n}} |\Delta|_1\end{aligned}$$

with probability at least $1 - 2\exp(-c' \log p)$. The first inequality follows from the triangle inequality and the last inequality from the definition of $\mathbf{h} = \mathrm{vec}(\Delta)$ and $|\cdot|_1$. □

7. Nonconvex Regularizers: Proof of Theorem 3

Recall that the support sets $\mathcal{S}, \mathcal{S}_k$ are the set of nonzero elements of Ω_0 and $\Psi_{k,0}$, respectively. Define $\mathcal{B}$ to be the set of matrices in $\mathcal{K}_\mathbf{p}$ with support contained in $\mathcal{S}$, that is

$$\mathcal{B} = \{\Omega = \Psi_1 \oplus \cdots \oplus \Psi_K \in \mathcal{K}_\mathbf{p} | \mathrm{supp}(\Psi_k) \subseteq \mathcal{S}_k, \, \forall k\}.$$

The set $\mathcal{B}$ is the set of Kronecker sum matrices following the true sparsity pattern of the Kronecker sum $\Omega_0 = \Psi_{1,0} \oplus \cdots \oplus \Psi_{K,0}$.

Note that $\mathcal{B}$ is a linear subspace of $\mathbb{R}^{p\times p}$ since $\mathcal{K}_\mathbf{p}$ is a linear subspace and the intersection of two linear subspaces is a linear subspace. Hence the (L2 norm) projection $\mathrm{Proj}_\mathcal{B} : \mathbb{R}^{p\times p} \to \mathcal{B}$ onto $\mathcal{B}$ is given by

$$\mathrm{Proj}_\mathcal{B}(A) = \mathrm{Proj}_\mathcal{S}(\mathrm{Proj}_{\mathcal{K}_\mathbf{p}}(A)),$$

where $\mathrm{Proj}_\mathcal{S}$ is the linear projection operator projecting $\mathbb{R}^{p\times p}$ onto matrices in $\mathbb{R}^{p\times p}$ with sparsity pattern $\mathcal{S}$, and $\mathrm{Proj}_{\mathcal{K}_\mathbf{p}}$ is the previously defined projection onto $\mathcal{K}_\mathbf{p}$ defined in Section 2 of the main text. Note that since the sparsity pattern $\mathcal{S}$ is the sparsity pattern of a Kronecker sum matrix in $\mathcal{K}_\mathbf{p}$, projection onto $\mathcal{S}$ does not change the Kronecker structure.

By reshaping we obtain the representation

$$\mathrm{vec}(\mathrm{Proj}_\mathcal{B}(A)) = \mathcal{P}_\mathcal{B}\mathrm{vec}(A) \tag{69}$$

where $\mathcal{P}_\mathcal{B} \in \mathbb{R}^{p^2\times p^2}$ is the *projection matrix* associated with the linear subspace $\mathcal{B}$. Recall that $\mathrm{vec}(\cdot)$ is the vectorization operator, and the *projection matrix* in linear algebra is UU^T where U is an orthonormal basis for the subspace.

We first summarize the proof of Theorem 3.

Proof plan: The proof concept is to apply the primal-dual witness technique of Loh et al. (2017) to our sparse Kronecker sum precision matrix estimator. Since the nonconvex graphical lasso proof in Loh et al. (2017) relied on the set of $\mathcal{S}$ sparse matrices being a linear subspace of $\mathbb{R}^{p\times p}$, we can simply replace the sparse subspace in their proof with our sparse Kronecker sum subspace $\mathcal{B}$ and proceed in a similar fashion. The primal-dual witness technique can be briefly summarized as

(i) Prove the regularized objective function (8) is strictly convex over the constraint set, so that any zero subgradient point is the unique global minimizer.

(ii) Construct a zero subgradient point of the *oracle* estimator objective function using Brouwer's theorem.
(iii) Prove this zero subgradient point $\widehat{\Omega}_{\text{oracle}}$ converges to the true Ω_0.
(iv) Prove that the zero subgradient point of the *oracle* objective is also a zero subgradient point of the full objective function (8), hence it is the unique global minimizer and converges to Ω_0.

Proceeding with the full proof, we first have the following lemma.

LEMMA 21. *Suppose g_ρ is μ-amenable. Then for $\kappa = \sqrt{\frac{2}{\mu}}$, the objective function* (8) *is strictly convex over the constraint set.*

PROOF. Recall that

$$\nabla^2 \left(-\log|\Omega| + \langle \widehat{S}, \Omega\rangle\right) = (\Omega \otimes \Omega)^{-1} \tag{70}$$

which is a deterministic quantity not depending on the data. Hence, for $\|\Omega\|_2 \leq \sqrt{1/\mu}$, the minimum eigenvalue satisfies

$$\lambda_{\min}(\nabla^2 \left(-\log|\Omega| + \langle \widehat{S}, \Omega\rangle\right)) = \lambda_{\min}((\Omega \otimes \Omega)^{-1}) \geq \frac{\mu}{2}.$$

This implies that $-\log|\Omega| + \langle \widehat{S}, \Omega\rangle - \frac{\mu}{2}\|\Omega\|_F^2$ is convex for $\|\Omega\|_2 \leq \sqrt{1/\mu}$. Furthermore, by μ-amenability, $\sum_{k=1}^K m_k \sum_{i\neq j} g_\lambda([\Psi_k]_{ij}) + \frac{\mu}{2}\|\Omega\|_F^2$ is convex for $\Omega \in \mathcal{K}_{\mathbf{p}}$. Therefore, since $\mathcal{K}_{\mathbf{p}}$ is a linear subspace, the complete objective (8) is convex for $\|\Omega\|_2 \leq \sqrt{1/\mu}$ and $\Omega \in \mathcal{K}_{\mathbf{p}}$. Since it is convex over $\mathcal{K}_{\mathbf{p}}$, it is convex over $\mathcal{K}_{\mathbf{p}}^\sharp$ as well, since $\mathcal{K}_{\mathbf{p}}^\sharp$ is the intersection of $\mathcal{K}_{\mathbf{p}}$ and the convex positive definite cone. □

Since the objective is convex, a point in the subspace $\mathcal{K}_{\mathbf{p}}$ with zero subgradient will be the unique global minimum. Our first step will be to construct such a zero subgradient point.

We will first construct the (unique) oracle estimate where the oracle gives the support set of Ω_0. We will then show that this oracle estimate is also a zero-subgradient point of the objective (8) and therefore its unique global minimizer.

Using the $\mathcal{B}$ notation, we can write the oracle estimate as

$$\widehat{\Omega}_{\text{oracle}} = \arg\min_{\Omega\in\mathcal{B}} -\log|\Omega| + \langle \widehat{S}, \Omega\rangle. \tag{71}$$

Our goal will be to construct a map $F : \mathcal{B} \to \mathcal{B}$ such that (a) Δ is a fixed point of F if and only if $\Omega_0 + \Delta$ is a fixed point of the oracle estimate (71), (b) F maps the intersection $\mathcal{B} \cap \mathbb{B}_\infty(r)$ of $\mathcal{B}$ and the radius-r ℓ_∞-ball centered at the origin to itself for some r, and (c) this r is such that $\Omega = \Omega_0 + \Delta \succ 0$, for all $\Delta \in \mathcal{B} \cap \mathbb{B}_\infty(r)$. Then by Brouwer's fixed point theorem we can show that F must have a fixed point Δ_* in that ball. By construction (a) above, this fixed point Δ_* will correspond to a fixed point $\Omega_0 + \Delta^*$ in the oracle estimator objective, hence the oracle estimate will have ℓ_∞-ball error less than r.

For F, we will choose a Newton method step (gradient step preconditioned by inverse Hessian). Denote the pseudoinverse of a matrix A as $A^\dagger$. We now write the map $F : \mathcal{B} \to \mathcal{B}$ given by

$$F(\Delta_S) := -\Gamma^\dagger \text{vec}\left(\text{Proj}_{\mathcal{B}}(\widehat{S} - (\Omega_0 + \Delta_S)^{-1})\right) + \text{vec}(\Delta_S)$$

where $\Delta_S \in \mathcal{B}$, and we let Γ be the Hessian of the objective function within $\mathcal{B}$:†

$$\Gamma = \mathcal{P}_{\mathcal{B}}(\Sigma_0 \otimes \Sigma_0)\mathcal{P}_{\mathcal{B}}.$$

The quantity $\Sigma_0 \otimes \Sigma_0$ is included as it is the Hessian of the objective function (70). The pseudoinverse is needed since $\mathcal{P}_{\mathcal{B}}$ is low rank, making the Hessian within $\mathcal{B}$ low rank.

Clearly if $\text{Proj}_{\mathcal{B}}(\widehat{S} - (\Omega_0 + \Delta_S)^{-1}) = 0$, $F(\Delta_S) = \Delta_S$ and vice versa, hence Δ_S is a fixed point of F if and only if $\Omega_0 + \Delta_S$ is a fixed point of the oracle objective (71). Now

$$\|\Delta_S\|_2 \leq dr$$

since Δ_S has at most d nonzero entries per row. Hence the matrix $\Omega_0 + \Delta_S$ is invertible and positive definite whenever $dr < \lambda_{\min}(\Omega_0)$, making F a continuous map on $\mathbb{B}_\infty(r) \cap \mathcal{B}$ and satisfying condition (c).

Define the constants $\kappa_\Gamma = \|\Gamma^\dagger\|_\infty$ and $\kappa_\Sigma = \|\Sigma_0\|_\infty$, in other words, we are assuming that the Hessian is well-conditioned in the ∞-norm sense, which is possible since Σ_0 has eigenvalues bounded from above and below. We now show the following lemma by verifying the remaining condition (b) on F and applying Brouwer's fixed point theorem. Several relevant quantities are summarized in Table 1 for convenience.

LEMMA 22. *Let $r = 2C_0\kappa_\Gamma\|\Sigma_0\|_2(K+1)\sqrt{\frac{\log p}{n \min_k m_k}}$ where C_0 is a constant depending only on the subgaussian parameter of the data and*

$$dr \leq \min\left\{\frac{1}{2}\lambda_{\min}(\Omega_0), \frac{1}{2\kappa_\Sigma}, \frac{1}{4\kappa_\Gamma\kappa_\Sigma^3}\right\}.$$

Assume the sample size satisfies $n \min_k m_k \geq \kappa_\Gamma^2 \log p$. Then under event $\mathcal{A}$ as in Theorem 1 there exists $\widehat{\Omega}_{\text{oracle}} \in \mathcal{B}$ such that

$$\|\widehat{\Omega}_{\text{oracle}} - \Omega_0\|_{\max} \leq r, \qquad \|\widehat{\Omega}_{\text{oracle}} - \Omega_0\|_2 \leq dr, \textit{and} \quad \text{Proj}_{\mathcal{B}}(\widehat{S} - \widehat{\Omega}_{\text{oracle}}^{-1}) = 0.$$

†With $\mathcal{P}_{\mathcal{B}} = UU^T$ as above (where columns of U form an orthonormal basis for the subspace $\mathcal{B}$), $\Gamma = UU^T(\Sigma_0 \otimes \Sigma_0)UU^T$ and hence $\Gamma^\dagger = U\left(U^T(\Sigma_0 \otimes \Sigma_0)U\right)^{-1}U^T$ since Σ_0 is positive definite.

Variable	Definition
$\mathcal{L}_n(\Omega)$	$-\log\lvert\Omega\rvert + \langle \widehat{S}, \Omega\rangle\Big\rvert_{\Omega\in\mathcal{K}_{\mathbf{p}}}$: Objective function less regularization terms.
$q_\rho(t)$	$g_\rho(t) - \rho\lvert t\rvert$: Difference between regularizer and $\ell 1$ penalty.
$\mathcal{K}_{\mathbf{p}}$	Set of Kronecker sum matrices with fixed dimensions $\mathbf{p} = [d_1, \dots, d_K]$.
$\mathcal{B}$	Set of matrices in $\mathcal{K}_{\mathbf{p}}$ with support contained in $\mathcal{S}$.
$\mathrm{Proj}_{\mathcal{B}}(\cdot)$	Linear projection operator from $\mathbb{R}^{p\times p}$ onto $\mathcal{B}$.
$\mathcal{P}_{\mathcal{B}}$	Projection matrix corresponding to $\mathrm{Proj}_{\mathcal{B}}$. $\mathcal{P}_{\mathcal{B}}\mathrm{vec}(A) = \mathrm{vec}(\mathrm{Proj}_{\mathcal{B}}(A))$.
Γ	$\mathcal{P}_{\mathcal{B}}(\Sigma_0 \otimes \Sigma_0)\mathcal{P}_{\mathcal{B}}$: Hessian of $\mathcal{L}_n$ within subspace $\mathcal{B}$.
κ_Γ	$\lVert\Gamma^\dagger\rVert_\infty$
κ_Σ	$\lVert\Sigma_0\rVert_\infty$
τ_Σ	$\frac{\mathrm{tr}(\widehat{S})-\mathrm{tr}(\Sigma_0)}{p}$
$\mathbf{e}_i$	ith unit vector in $\mathbb{R}^p$.

Table 1: Selected quantities used in the proof of Theorem 3

PROOF. First, note that $\Gamma^\dagger\Gamma\mathrm{vec}(\Delta) = \mathrm{vec}(\Delta)$ for any $\Delta \in \mathcal{B}$, since Γ is the projection of the positive definite matrix $\Sigma_0 \otimes \Sigma_0$ onto the low rank subspace $\mathcal{B}$.

Suppose $\Delta_S \in \mathbb{B}_\infty(r)$. Then

$$
\begin{aligned}
F(\mathrm{vec}(\Delta_S)) := -\Gamma^\dagger &\left\{\mathrm{vec}\left(\mathrm{Proj}_{\mathcal{B}}(\widehat{S} - \Sigma_0)\right)\right.\\
&\left.+\mathrm{vec}\left(\mathrm{Proj}_{\mathcal{B}}(\Sigma_0 - (\Omega_0 + \Delta_S)^{-1})\right) + \Gamma\mathrm{vec}(\Delta_S)\right\},
\end{aligned}
$$

hence

$$
\begin{aligned}
\lVert F(\mathrm{vec}(\Delta_S))\rVert_\infty \leq &\kappa_\Gamma \lVert\mathrm{vec}\left(\mathrm{Proj}_{\mathcal{B}}(\widehat{S} - \Sigma_0)\right)\rVert_\infty \qquad (72)\\
&+ \kappa_\Gamma\lVert\mathrm{vec}\left(\mathrm{Proj}_{\mathcal{B}}(\Sigma_0 - (\Omega_0 + \Delta_S)^{-1})\right) + \Gamma\mathrm{vec}(\Delta_S)\rVert_\infty,
\end{aligned}
$$

by the definition of κ_Γ and the triangle inequality.

The first term of (72) can be bounded via the concentration inequalities used for the $\ell 1$ case. Specifically, note that

$$
\begin{aligned}
&\kappa_\Gamma\lVert\mathrm{vec}\left(\mathrm{Proj}_{\mathcal{B}}(\widehat{S} - \Sigma_0)\right)\rVert_\infty \qquad (73)\\
&\leq \kappa_\Gamma\lVert\mathrm{vec}\left(\mathrm{Proj}_{\mathcal{K}_{\mathbf{p}}}(\widehat{S} - \Sigma_0)\right)\rVert_\infty\\
&= \kappa_\Gamma\lVert(S_1 - \Sigma_0^{(1)} \oplus \dots \oplus (S_K - \Sigma_0^{(K)}) - \tau_\Sigma I_p\rVert_{\max}\\
&\leq \kappa_\Gamma \sum_{k=1}^{K} \lVert S_k - \Sigma_0^{(k)}\rVert_{\max} + \frac{\lvert\mathrm{tr}(\widehat{S} - \mathrm{tr}(\Sigma_0)\rvert}{p}.
\end{aligned}
$$

where we have used $\tau_\Sigma = \frac{\mathrm{tr}(\widehat{S})-\mathrm{tr}(\Sigma_0)}{p}$. Now recall that under event $\mathcal{A}_k$, defined in (47) above,

$$\max_{ij}\left|[S_k - \Sigma_0^{(k)}]_{ij}\right| = \max_{ij}\left|\langle \mathbf{e}_i\mathbf{e}_j^T, S_k - \Sigma_0^{(k)}\rangle\right| \le C_0\|\Sigma_0\|_2\sqrt{\frac{\log p}{m_k n}},$$

and under event $\mathcal{A}_0$, defined above in (46),

$$|\tau_\Sigma| \le C_0\|\Sigma_0\|_2\sqrt{\frac{\log p}{pn}}.$$

Hence under event $\mathcal{A} = \bigcup_{k=0}^K \mathcal{A}_k$,

$$\begin{aligned}\kappa_\Gamma\|\mathrm{vec}\left(\mathrm{Proj}_{\mathcal{B}}(\widehat{S} - \Sigma_0)\right)\|_\infty &\le C_0\kappa_\Gamma\|\Sigma_0\|_2\left[\sqrt{\frac{\log p}{pn}} + \sum_{k=1}^K\sqrt{\frac{\log p}{m_k n}}\right]\\ &\le C_0\kappa_\Gamma\|\Sigma_0\|_2(K+1)\sqrt{\frac{\log p}{n\min_k m_k}}\\ &= \frac{r}{2}. \qquad (74)\end{aligned}$$

Finally, recall that by Lemma 11 event $\mathcal{A}$ holds with probability $\ge 1 - 2(K+1)\exp(-c\log p)$.

Moving on to the second term of (72), we apply the matrix expansion

$$(A+\Delta)^{-1} - A^{-1} = \sum_{\ell=1}^\infty (-A^{-1}\Delta)^\ell A^{-1}, \qquad (75)$$

and note that (since $\Delta_S \in \mathcal{B}$ implies $\mathcal{P}_{\mathcal{B}}\mathrm{vec}(\Delta_S) = \mathrm{vec}(\Delta_S)$)

$$\begin{aligned}\Gamma\mathrm{vec}(\Delta_S) &= \mathcal{P}_{\mathcal{B}}\left((\Sigma_0\otimes\Sigma_0)\mathcal{P}_{\mathcal{B}}\mathrm{vec}(\Delta_S)\right)\\ &= \mathrm{vec}\left(\mathrm{Proj}_{\mathcal{B}}\left(\mathrm{vec}^{-1}\left((\Sigma_0\otimes\Sigma_0)\mathrm{vec}(\Delta_S)\right)\right)\right)\\ &= \mathrm{vec}(\mathrm{Proj}_{\mathcal{B}}(\Sigma_0\Delta_S\Sigma_0)),\end{aligned}$$

where we have used the fact that for symmetric matrices A, B, $\mathrm{vec}(ABA) = (A\otimes A)\mathrm{vec}(B)$.

We then obtain

$$\begin{aligned}&\mathrm{vec}\left(\mathrm{Proj}_{\mathcal{B}}(\Sigma_0 - (\Omega_0+\Delta_S)^{-1})\right) + \Gamma\mathrm{vec}(\Delta_S)\\ &= \mathrm{vec}\left(\mathrm{Proj}_{\mathcal{B}}(\Sigma_0 - (\Omega_0+\Delta_S)^{-1} - \Sigma_0\Delta_S\Sigma_0)\right)\\ &= \mathrm{vec}\left(\mathrm{Proj}_{\mathcal{B}}\left(\sum_{\ell=2}^\infty(-\Sigma_0\Delta_S)^\ell\Sigma_0\right)\right).\end{aligned}$$

We have used $\mathrm{vec}^{-1}(\cdot)$ to denote the inverse of the vectorization operator.

Via the triangle inequality and the linearity of the vectorization and projection operators,

$$\begin{aligned}&\|\mathrm{vec}\left(\mathrm{Proj}_{\mathcal{B}}(\Sigma_0 - (\Omega_0 + \Delta_S)^{-1})\right) + \Gamma \mathrm{vec}(\Delta_S)\|_\infty \\ &\leq \max_{(j,k)\in S} \sum_{\ell=2}^{\infty} |\mathbf{e}_j^T (\Sigma_0 \Delta)^\ell \Sigma_0 \mathbf{e}_k|. \end{aligned} \tag{76}$$

Now we can apply Holder's inequality to obtain

$$\begin{aligned}|\mathbf{e}_j^T(\Sigma_0\Delta)^\ell \Sigma_0 \mathbf{e}_k| &\leq \|\mathbf{e}_j^T(\Sigma_0\Delta)^{\ell-1}\Sigma_0\|_1 \|\Delta\Sigma_0\|_\infty \\ &\leq \|\Sigma_0(\Delta\Sigma_0)^{\ell-1}\|_1 \|\Delta\|_{\max} \|\Sigma_0 e_k\|_1 \\ &\leq \|\Sigma_0\|_1^{\ell-1}\|\Delta\|_{\max}\|\Sigma_0\|_1 \\ &= \|\Sigma_0\|_\infty^{\ell+1}\|\Delta\|_\infty^{\ell-1}\|\Delta\|_{\max}.\end{aligned}$$

Then, using the fact that $\|\Delta\|_2 \leq \|\Delta\|_\infty \leq dr$ and substituting back into (76), we have

$$\begin{aligned}&\|\mathrm{vec}\left(\mathrm{Proj}_{\mathcal{B}}(\Sigma_0 - (\Omega_0 + \Delta_S)^{-1})\right) + \Gamma \mathrm{vec}(\Delta_S)\|_\infty \\ &\leq \sum_{\ell=2}^{\infty} \kappa_\Sigma^{\ell+1} d^{\ell-2} r^\ell \\ &= \frac{\kappa_\Sigma^3 dr^2}{1-\kappa_\Sigma dr} \\ &\leq 2\kappa_\Sigma^3 dr^2.\end{aligned}$$

Since our assumption implies that $2\kappa_\Sigma^3 dr^2 \leq r$, we therefore have that

$$\|F(\mathrm{vec}(\Delta_S))\|_\infty \leq r$$

under event $\mathcal{A}$. Since $F(\mathbb{B}_\infty(r) \cap \mathcal{B}) \in \mathbb{B}_\infty(r) \cap \mathcal{B}$, by Brouwer's fixed point theorem (Ortega and Rheinboldt, 1970), F must have a fixed point Δ_S^*. Recalling that $\Delta_S^*, \Omega_0 \in \mathcal{B}$, we choose $\widehat{\Omega}_{\text{oracle}} = \Omega_0 + \Delta_S^*$. Hence by construction $\|\widehat{\Omega}_{\text{oracle}} - \Omega_0\|_{\max} \leq r$ and $\|\widehat{\Omega}_{\text{oracle}} - \Omega_0\|_2 \leq dr$ since both matrices have degree bounded by d. The last equality follows since Δ_S^* is the fixed point of F, i.e. $F(\Delta_S^*) = \Delta_S^*$, which can only occur if

$$\mathrm{vec}\left(\mathrm{Proj}_{\mathcal{B}}(\widehat{S} - (\Omega_0 + \Delta_S^*)^{-1})\right) = 0.$$

□

Using this lemma it remains to show that $\widehat{\Omega}_{\text{oracle}}$ satisfies the constraints and is a zero-subgradient point of the complete objective (8), and hence is the unique global optimum. Define $\mathcal{L}_n(\Omega)$ to be the objective function (8) less the regularization terms, i.e.

$$\mathcal{L}_n(\Omega) = -\log|\Omega| + \langle \widehat{S}, \Omega\rangle\Big|_{\Omega\in\mathcal{K}_{\mathbf{p}}}.$$

LEMMA 23. *The oracle estimate* $\widehat{\Omega}_{\text{oracle}}$ *will be a zero-subgradient point of the global objective* (8) *if the inequalities*

$$\|\nabla_{\mathcal{K}_{\mathbf{p}}}\mathcal{L}_n(\Omega_0)\|_\infty \le \frac{1}{2}\rho \tag{77}$$

and

$$\|\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^\dagger \nabla_{\mathcal{K}_{\mathbf{p}}}\mathcal{L}_n(\Omega_0)_S\|_\infty \le \frac{1}{2}\rho \tag{78}$$

hold, where

$$\widehat{Q} = \int_0^1 \nabla^2_{\mathcal{K}_{\mathbf{p}}}\mathcal{L}_n\left(\Omega_0 + t(\widehat{\Omega}_{\text{oracle}} - \Omega_0)\right) dt.$$

We have denoted $\nabla_{\mathcal{K}_{\mathbf{p}}} f = \mathcal{P}_{\mathcal{K}_{\mathbf{p}}}\nabla f$ *and* $\nabla^2_{\mathcal{K}_{\mathbf{p}}} f = \mathcal{P}_{\mathcal{K}_{\mathbf{p}}}(\nabla^2 f)\mathcal{P}_{\mathcal{K}_{\mathbf{p}}}$ *to be the gradient and Hessian respectively of* f *projected onto the subspace* $\mathcal{K}_{\mathbf{p}}$ *(*$\mathcal{P}_{\mathcal{K}_{\mathbf{p}}}$ *is the projection matrix onto* $\mathcal{K}_{\mathbf{p}}$*).*

PROOF. In this proof, for simplicity we write $q_\rho(\widehat{\Omega})$ to indicate $q_\rho(t) = g_\rho(t) - \rho|t|$ applied elementwise to the offdiagonal elements of $\widehat{\Omega}$:

$$[q_\rho(\widehat{\Omega})]_{ij} = \begin{cases} q_\rho(\widehat{\Omega}_{ij}) & i \neq j \\ 0 & \text{otherwise.} \end{cases}$$

Observe that by construction $\nabla_{\mathcal{K}_{\mathbf{p}}} q_\rho(\Omega) = \nabla_{\mathbb{R}^{p\times p}} q_\rho(\Omega) = \nabla q_\rho(\Omega)$ for any $\Omega \in \mathcal{K}_{\mathbf{p}}$.

For the objective (8), the zero subgradient condition is given by

$$\nabla_{\mathcal{K}_{\mathbf{p}}}\mathcal{L}_n(\widehat{\Omega}) - \nabla q_\rho(\widehat{\Omega}) + \rho\widehat{z} = 0,$$

where $\widehat{z} = \partial|\widehat{\Omega}|_{1,\text{off}}$ is an element of the subgradient of the off-diagonal $\ell 1$ norm at $\widehat{\Omega}$. Adding and subtracting $\nabla_{\mathcal{K}_{\mathbf{p}}}\mathcal{L}_n(\Omega_0)$ gives

$$(\nabla_{\mathcal{K}_{\mathbf{p}}}\mathcal{L}_n(\widehat{\Omega}) - \nabla_{\mathcal{K}_{\mathbf{p}}}\mathcal{L}_n(\Omega_0)) + (\nabla_{\mathcal{K}_{\mathbf{p}}}\mathcal{L}_n(\Omega_0) - \nabla q_\rho(\widehat{\Omega})) + \rho\widehat{z} = 0.$$

By the fundamental theorem of calculus we have (for $\widehat{\Omega} = \widehat{\Omega}_{\text{oracle}}$) that $\nabla_{\mathcal{K}_{\mathbf{p}}}\mathcal{L}_n(\widehat{\Omega}_{\text{oracle}}) - \nabla_{\mathcal{K}_{\mathbf{p}}}\mathcal{L}_n(\Omega_0) = \widehat{Q}\text{vec}(\widehat{\Omega}_{\text{oracle}} - \Omega_0)$, hence

$$\widehat{Q}\text{vec}(\widehat{\Omega}_{\text{oracle}} - \Omega_0) + (\nabla_{\mathcal{K}_{\mathbf{p}}}\mathcal{L}_n(\Omega_0) - \nabla q_\rho(\widehat{\Omega}_{\text{oracle}})) + \rho\widehat{z} = 0.$$

Rewriting in block form gives

$$\begin{bmatrix} \widehat{Q}_{SS} & \widehat{Q}_{SS^c} \\ \widehat{Q}_{S^cS} & \widehat{Q}_{S^cS^c} \end{bmatrix} \left(\widehat{\Omega}_{\mathrm{oracle}} - \Omega_0\right) \\ + \left(\begin{bmatrix} \nabla_{\mathcal{K}_{\mathbf{p}}} \mathcal{L}_n(\Omega_0)_S - \nabla q_\rho(\widehat{\Omega}_{\mathrm{oracle}})_S \\ \nabla_{\mathcal{K}_{\mathbf{p}}} \mathcal{L}_n(\Omega_0)_{S^c} - \nabla q_\rho(0) \end{bmatrix}\right) + \rho \begin{bmatrix} \widehat{z}_S \\ \widehat{z}_{S^c} \end{bmatrix} = 0,$$

where $\widehat{Q}_{SS}$ is the block of $\widehat{Q}$ corresponding to the elements in $\mathcal{S}$ along both axes, $\widehat{Q}_{S^cS^c}$ is the block of $\widehat{Q}$ corresponding to the elements in the complement of $\mathcal{S}$, etc. After some algebra we obtain a solution

$$\begin{aligned} \widehat{z}_{S^c} =& \frac{1}{\rho} \left\{ \left(\nabla q_\rho(0) - \nabla_{\mathcal{K}_{\mathbf{p}}} \mathcal{L}_n(\Omega_0)_{S^c}\right) \right. \\ & \left. + \widehat{Q}_{S^cS} \widehat{Q}_{SS}^{\dagger} \left(\nabla_{\mathcal{K}_{\mathbf{p}}} \mathcal{L}_n(\Omega_0)_S - \nabla q_\rho(\widehat{\Omega}_{\mathrm{oracle}})_S + \rho \widehat{z}_S\right) \right\}, \end{aligned}$$

since $\nabla q_\rho(0) = 0$ by definition. Now from Lemma 22, under event $\mathcal{A}$

$$\|\widehat{\Omega}_{\mathrm{oracle}} - \Omega_0\|_{\max} \le r,$$

and observe that $\rho\gamma > r$ since we have assumed that $n \min_k m_k \ge c_0 d^2 \log p$ for some c_0 large enough. By our assumption that $|[\Omega_0]_{ij}| \ge \rho\gamma + r$ for all i, j, we then have (again under event $\mathcal{A}$)

$$\min_{ij} |[\widehat{\Omega}_{\mathrm{oracle}}]_{ij}| \ge \rho\gamma + r - r = \rho\gamma.$$

Therefore, using condition (f) of the definition of a (μ, γ) regularizer, $-\nabla q_\rho(\widehat{\Omega}_{\mathrm{oracle}})_S + \rho \widehat{z}_S = 0$ and

$$\begin{aligned} \|\widehat{z}_{S^c}\|_\infty &= \frac{1}{\rho} \left\| -\nabla_{\mathcal{K}_{\mathbf{p}}} \mathcal{L}_n(\Omega_0)_{S^c} + \widehat{Q}_{S^cS} \widehat{Q}_{SS}^{\dagger} \nabla_{\mathcal{K}_{\mathbf{p}}} \mathcal{L}_n(\Omega_0)_S \right\|_\infty \\ &\le \frac{1}{\rho} \|\nabla_{\mathcal{K}_{\mathbf{p}}} \mathcal{L}_n(\Omega_0)_{S^c}\|_\infty + \frac{1}{\rho} \|\widehat{Q}_{S^cS} \widehat{Q}_{SS}^{\dagger} \nabla_{\mathcal{K}_{\mathbf{p}}} \mathcal{L}_n(\Omega_0)_S\|_\infty \\ &\le \frac{1}{\rho} \|\nabla_{\mathcal{K}_{\mathbf{p}}} \mathcal{L}_n(\Omega_0)\|_\infty + \frac{1}{\rho} \|\widehat{Q}_{S^cS} \widehat{Q}_{SS}^{\dagger} \nabla_{\mathcal{K}_{\mathbf{p}}} \mathcal{L}_n(\Omega_0)_S\|_\infty \\ &\le \frac{1}{2} + \frac{1}{2} = 1, \end{aligned}$$

where we have applied the assumed inequalities. Since $\|\widehat{z}_{S^c}\|_\infty \le 1$, it is a feasible subgradient and therefore $\widehat{\Omega}_{\mathrm{oracle}}$ is a zero subgradient point of the global objective function (8).

□

We now show the inequalities (77), (78) assumed by Lemma 23 hold under event $\mathcal{A}$. Note that

$$\|\nabla_{\mathcal{K}_{\mathbf{p}}} \mathcal{L}_n(\Omega_0)\|_\infty = \|\mathrm{Proj}_{\mathcal{K}_{\mathbf{p}}}(\widehat{S} - \Sigma_0)\|_{\max}$$

and thus by (74), under event $\mathcal{A}$ equation (77) holds with $\rho = \frac{r}{\kappa_\Gamma}$.

It remains to show (78) holds with $\rho = r$. We will first bound

$$\|(\nabla^2_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0))_{S^cS}(\nabla^2_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0))^\dagger_{SS}(\nabla_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0)_S\|_\infty,$$

and then show that the expression on the left hand side of (78) is close to this quantity.

First, by the definition of the infinity norm and $\mathcal{P}_{\mathcal{K}_\mathbf{p}}$ it can be shown that

$$\|\mathcal{P}_{\mathcal{K}_\mathbf{p}}\|_\infty = \sup_{A\in\mathbb{R}^{p\times p}} \frac{\|\mathcal{P}_{\mathcal{K}_\mathbf{p}}\mathrm{vec}(A)\|_\infty}{\|A\|_{\max}} = \sup_{A\in\mathbb{R}^{p\times p}} \frac{\|\mathrm{Proj}_{\mathcal{K}_\mathbf{p}}(A)\|_{\max}}{\|A\|_{\max}} \leq 2K, \tag{79}$$

where we have used the expression (95) for the elements of the projected matrix and the fact that an average of a set of elements of A cannot have magnitude larger than $\|A\|_{\max}$. Noting that $(\nabla^2_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0))^\dagger_{SS} = (\Gamma^\dagger)_{SS}$,

$$\begin{aligned}&\|(\nabla^2_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0))_{S^cS}(\nabla^2_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0))^\dagger_{SS}(\nabla_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0)_S\|_\infty \\ &\quad\leq \|(\nabla^2_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0))_{S^cS}\|_\infty \cdot \|(\Gamma^\dagger)_{SS}\|_\infty \cdot \|(\nabla_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0)_S\|_\infty \\ &\quad\leq O\left(\frac{r}{\kappa_\Gamma}\right),\end{aligned} \tag{80}$$

since $\|(\nabla^2_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0))_{S^cS}\|_\infty \leq \|\nabla^2_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0)\|_\infty = \left\|\mathcal{P}_{\mathcal{K}_\mathbf{p}}\left(\nabla^2\mathcal{L}_n(\Omega_0)\right)\mathcal{P}_{\mathcal{K}_\mathbf{p}}\right\|_\infty \leq \|\mathcal{P}_{\mathcal{K}_\mathbf{p}}\|^2_\infty\|\nabla^2\mathcal{L}_n(\Omega_0)\|_\infty = \|\mathcal{P}_{\mathcal{K}_\mathbf{p}}\|^2_\infty\|\Sigma_0\otimes\Sigma_0\|_\infty = \|\mathcal{P}_{\mathcal{K}_\mathbf{p}}\|^2_\infty\|\Sigma_0\|^2_\infty \leq 4K^2\kappa^2_\Sigma$, and (77) holds under event $\mathcal{A}$ with $\rho = \frac{r}{\kappa_\Gamma}$.

We now relate the bound in (80) to that required to show (78). Note that

$$\begin{aligned}&\|\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^\dagger\nabla_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0)_S\|_\infty \\ &\leq \|(\nabla^2_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0))_{S^cS}(\nabla^2_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0))^\dagger_{SS}(\nabla_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0)_S)\|_\infty + \|\Xi(\nabla_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0)_S)\|_\infty \\ &\leq O\left(\frac{r}{\kappa_\Gamma}\right) + \|\Xi(\nabla_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0)_S)\|_\infty,\end{aligned} \tag{81}$$

where we have defined

$$\Xi = \widehat{Q}_{S^cS}(\widehat{Q}_{SS})^\dagger - (\nabla^2_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0))_{S^cS}(\nabla^2_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0))^\dagger_{SS}.$$

Now, again invoking (77),

$$\|\Xi(\nabla_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0)_S)\|_\infty \leq \|\Xi\|_\infty \cdot \|\nabla_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0)_S\|_\infty \leq \|\Xi\|_\infty O\left(r\right). \tag{82}$$

The infinity norm of Ξ can be bounded as

$$\begin{aligned}\|\Xi\|_\infty \leq &\left\|\left(\widehat{Q}_{S^cS} - (\nabla^2_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0))_{S^cS}\right)\left((\widehat{Q}_{SS})^\dagger - (\nabla^2_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0))^\dagger_{SS}\right)\right\|_\infty \qquad (83)\\ &+ \left\|(\widehat{Q}_{S^cS} - (\nabla^2_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0))_{S^cS})(\nabla^2_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0))^\dagger_{SS}\right\|_\infty \\ &+ \left\|(\nabla^2_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0))_{S^cS}\left((\widehat{Q}_{SS})^\dagger - (\nabla^2_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0))^\dagger_{SS}\right)\right\| \\ &\leq \delta_1\delta_2 + \delta_1\|(\widehat{Q}_{SS})^\dagger\|_\infty + \delta_2\|(\nabla^2_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0))_{S^cS}\|_\infty\end{aligned}$$

where we have set

$$\begin{aligned}\delta_1 &:= \|\widehat{Q}_{S^cS} - (\nabla^2_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0))_{S^cS}\|_\infty\\ \delta_2 &:= \|(\widehat{Q}_{SS})^\dagger - (\nabla^2_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0))^\dagger_{SS}\|_\infty.\end{aligned}$$

First note that by (79)

$$\begin{aligned}\|(\nabla^2_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0))_{S^cS}\|_\infty &\leq \|(\nabla^2_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0))\|_\infty\\ &\leq \|\mathcal{P}_{\mathcal{K}_\mathbf{p}}(\nabla^2\mathcal{L}_n(\Omega_0))\mathcal{P}_{\mathcal{K}_\mathbf{p}}\|_\infty\\ &\leq \|\mathcal{P}_{\mathcal{K}_\mathbf{p}}\|^2_\infty\|\nabla^2\mathcal{L}_n(\Omega_0)\|_\infty\\ &\leq 4K^2\|\nabla^2\mathcal{L}_n(\Omega_0)\|_\infty \leq 4K^2\kappa^2_\Sigma,\end{aligned}$$

and $\|(\widehat{Q}_{SS})^\dagger\|_\infty = O(1+\delta_2)$ by the definition of δ_2.

Substituting into (83) gives

$$\|\Xi\|_\infty \leq O(\delta_1\delta_2) + O(\delta_1) + O(\delta_2). \qquad (84)$$

We bound δ_1 and δ_2 in the following lemma, proved in Section 7.1.

LEMMA 24. *Under the conditions of Lemma 22,*

$$\begin{aligned}\delta_1 &= O(dr)\\ \delta_2 &= O(dr).\end{aligned}$$

Applying Lemma 24 to (84), we obtain

$$\|\Xi\|_\infty = O(d^2r^2) + O(dr) + O(dr) = O(dr)$$

and substituting into (82)

$$\|\Xi(\nabla_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0)_S)\|_\infty = O(dr)\cdot O(r) = O(r),$$

since $dr = o(1)$ by our assumption that $n\min_k m_k \geq c_0 d^2\log p$ for some c_0 large enough.

Therefore, substituting into (81) we obtain

$$\|\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^\dagger \nabla_{\mathcal{K}_{\mathrm{p}}}\mathcal{L}_n(\Omega_0)_S\|_\infty = O(r) + O(r) = O(r),$$

proving the desired condition (78) holds. Hence the conditions of Lemma 23 hold under event $\mathcal{A}$, and $\widehat{\Omega}_{\mathrm{oracle}}$ is the unique global minimizer of the complete objective (8).

The Frobenius and spectral norm bounds follow from the identities

$$\|\widehat{\Omega} - \Omega_0\|_F \leq \sqrt{s+p}\|\widehat{\Omega} - \Omega_0\|_{\max}$$

and

$$\|\widehat{\Omega} - \Omega_0\|_2 \leq d\|\widehat{\Omega} - \Omega_0\|_{\max},$$

where the latter identity follows by symmetry of Ω.

7.1. Proof of Lemma 24: Bound on δ_1, δ_2

PROOF. Consider that

$$\begin{aligned}\widehat{Q} - \nabla^2_{\mathcal{K}_{\mathrm{p}}}\mathcal{L}_n(\Omega_0) &= \int_0^1 \nabla^2_{\mathcal{K}_{\mathrm{p}}}\mathcal{L}_n\left(\Omega_0 + t(\widehat{\Omega}_{\mathrm{oracle}} - \Omega_0)\right) dt - \nabla^2_{\mathcal{K}_{\mathrm{p}}}\mathcal{L}_n(\Omega_0)\\ &= \mathcal{P}_{\mathcal{K}_{\mathrm{p}}}\left(\int_0^1 \nabla^2\mathcal{L}_n\left(\Omega_0 + t(\widehat{\Omega}_{\mathrm{oracle}} - \Omega_0)\right) dt - \nabla^2\mathcal{L}_n(\Omega_0)\right)\mathcal{P}_{\mathcal{K}_{\mathrm{p}}}.\end{aligned}$$

Hence, since $\|\mathcal{P}_{\mathcal{K}_{\mathrm{p}}}\|_\infty \leq 2K$ by (79),

$$\begin{aligned}&\|\widehat{Q} - \nabla^2_{\mathcal{K}_{\mathrm{p}}}\mathcal{L}_n(\Omega_0)\|_\infty\\ &\leq \|\mathcal{P}_{\mathcal{K}_{\mathrm{p}}}\|_\infty^2 \left\|\int_0^1 \nabla^2\mathcal{L}_n\left(\Omega_0 + t(\widehat{\Omega}_{\mathrm{oracle}} - \Omega_0)\right) dt - \nabla^2\mathcal{L}_n(\Omega_0)\right\|_\infty\\ &\leq 4K^2\left\|\int_0^1 (\Omega_0 + t(\widehat{\Omega}_{\mathrm{oracle}} - \Omega_0))^{-1} \otimes (\Omega_0 + t(\widehat{\Omega}_{\mathrm{oracle}} - \Omega_0))^{-1} - \Omega_0^{-1}\otimes\Omega_0^{-1} dt\right\|_\infty\\ &\leq 4K^2\int_0^1 \left\|(\Omega_0 + t(\widehat{\Omega}_{\mathrm{oracle}} - \Omega_0))^{-1} \otimes (\Omega_0 + t(\widehat{\Omega}_{\mathrm{oracle}} - \Omega_0))^{-1} - \Omega_0^{-1}\otimes\Omega_0^{-1}\right\|_\infty dt.\end{aligned}$$

By Lemma 22, for $t \in [0,1]$,

$$\|\Omega_0 + t(\widehat{\Omega}_{\mathrm{oracle}} - \Omega_0) - \Omega_0\|_\infty = t\|\widehat{\Omega}_{\mathrm{oracle}} - \Omega_0\|_\infty \leq d\|\widehat{\Omega}_{\mathrm{oracle}} - \Omega_0\|_{\max} \leq dr.$$

We make use of the following matrix inequalities (Loh et al., 2017). For any invertible $A, B \in \mathbb{R}^{p\times p}$ and matrix norm $\|\cdot\|$,

$$\|A^{-1} - B^{-1}\| \leq \frac{\|A^{-1}\|^2\|A-B\|}{1 - \|A^{-1}\|\|A-B\|} = O(\|A^{-1}\|^2\|A-B\|). \tag{85}$$

if $\|A^{-1}\| \|A - B\| \leq 1/2$. For any A and B matrices of equal dimension we have

$$\|A \otimes A - B \otimes B\|_\infty \leq \|A - B\|_\infty^2 + 2 \min(\|A\|_\infty, \|B\|_\infty) \cdot \|A - B\|_\infty. \tag{86}$$

Applying (85) we get

$$\begin{aligned}
&\left\| \left(\Omega_0 + t(\widehat{\Omega}_{\text{oracle}} - \Omega_0)\right)^{-1} - \Omega_0^{-1} \right\|_\infty \\
&\leq O\left(\left\|\Omega_0^{-1}\right\|_\infty^2 \|\Omega_0 + t(\widehat{\Omega}_{\text{oracle}} - \Omega_0) - \Omega_0\|_\infty \right) \\
&\leq O(dr),
\end{aligned}$$

since $\|\Omega_0^{-1}\|_\infty = \|\Sigma_0\|_\infty$ is bounded by κ_Σ. Applying (86) to this yields

$$\|\widehat{Q} - \nabla^2_{\mathcal{K}_\mathbf{p}} \mathcal{L}_n(\Omega_0)\|_\infty = O(dr),$$

which gives

$$\delta_1 = \|\widehat{Q}_{S^c S} - (\nabla^2_{\mathcal{K}_\mathbf{p}} \mathcal{L}_n(\Omega_0))_{S^c S}\|_\infty = O(dr)$$

and

$$\|\widehat{Q}_{SS} - (\nabla^2_{\mathcal{K}_\mathbf{p}} \mathcal{L}_n(\Omega_0))_{SS}\|_\infty = O(dr). \tag{87}$$

Finally, recall that the projection matrix onto $\mathcal{K}_\mathbf{p}$ can be written as UU^T with $U^T U = I$ so

$$\begin{aligned}
\delta_2 &= \|(\widehat{Q}_{SS})^\dagger - (\nabla^2_{\mathcal{K}_\mathbf{p}} \mathcal{L}_n(\Omega_0))^\dagger_{SS}\|_\infty \\
&= \left\| U \left(\left(U^T \widehat{Q}_{SS} U \right)^{-1} - \left(U^T \nabla^2_{\mathcal{K}_\mathbf{p}} \mathcal{L}_n(\Omega_0)_{SS} U \right)^{-1} \right) U^T \right\|_\infty .
\end{aligned}$$

By the matrix expansion (75) we then have

$$\begin{aligned}
&(\widehat{Q}_{SS})^\dagger - (\nabla^2_{\mathcal{K}_\mathbf{p}} \mathcal{L}_n(\Omega_0))^\dagger_{SS} \\
&= U \left[\left(U^T \widehat{Q}_{SS} U \right)^{-1} - \left(U^T \nabla^2_{\mathcal{K}_\mathbf{p}} \mathcal{L}_n(\Omega_0)_{SS} U \right)^{-1} \right] U^T \\
&= U \left[\sum_{\ell=1}^\infty \left[\left(- \left(U^T \nabla^2_{\mathcal{K}_\mathbf{p}} \mathcal{L}_n(\Omega_0)_{SS} U \right)^{-1} \left(U^T \widehat{Q}_{SS} U - U^T \nabla^2_{\mathcal{K}_\mathbf{p}} \mathcal{L}_n(\Omega_0)_{SS} U \right) \right)^\ell \right.\right. \\
&\qquad\qquad \left.\left. \cdot \left(U^T \nabla^2_{\mathcal{K}_\mathbf{p}} \mathcal{L}_n(\Omega_0)_{SS} U \right)^{-1} \right] \right] U^T \\
&= \sum_{\ell=1}^\infty \left(- \left((\nabla^2_{\mathcal{K}_\mathbf{p}} \mathcal{L}_n(\Omega_0))^\dagger_{SS} \right) \left(\widehat{Q}_{SS} - (\nabla^2_{\mathcal{K}_\mathbf{p}} \mathcal{L}_n(\Omega_0))_{SS} \right) \right)^\ell \left((\nabla^2_{\mathcal{K}_\mathbf{p}} \mathcal{L}_n(\Omega_0))^\dagger_{SS} \right).
\end{aligned}$$

We can then use the bound (87) to obtain

$$\begin{aligned}\delta_2 &= \|(\widehat{Q}_{SS})^\dagger - (\nabla^2_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0))^\dagger_{SS}\|_\infty \\ &\leq \sum_{\ell=1}^{\infty} O\left(\|\widehat{Q}_{SS} - (\nabla^2_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0))_{SS}\|^\ell_\infty\right) \\ &= O\left(\frac{\|\widehat{Q}_{SS} - (\nabla^2_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0))_{SS}\|_\infty}{1 - \|\widehat{Q}_{SS} - (\nabla^2_{\mathcal{K}_\mathbf{p}}\mathcal{L}_n(\Omega_0))_{SS}\|_\infty}\right) \\ &= O(dr),\end{aligned}$$

since $dr = o(1)$.

□

8. Numerical Convergence of TG-ISTA

The following theorem shows that the iterates of the TG-ISTA implementation of TeraLasso converge geometrically to the global minimum:

THEOREM 25. *Let $\rho_k \geq 0$ for all k and let Ω_{init} be the initialization of the TG-ISTA implementation of TeraLasso (Algorithm 4). Let*

$$a = \frac{1}{\sum_{k=1}^{K} \|S_k\|_2 + d_k\rho_k}, \qquad b = \|\Omega^*\|_2 + \|\Omega_{\text{init}} - \Omega^*\|_F,$$

and assume $\zeta_t \leq a^2$ for all t. Suppose further that Ω^* is the global optimum. Then

$$\|\Omega_{t+1} - \Omega^*\|_F \leq \max\left\{\left|1 - \frac{\zeta_t}{b^2}\right|, \left|1 - \frac{\zeta_t}{a^2}\right|\right\} \|\Omega_t - \Omega^*\|_F.$$

Furthermore, the step size ζ_t which yields an optimal worst-case contraction bound $s(\zeta_t)$ is $\zeta = \frac{2}{a^{-2}+b^{-2}}$. The corresponding optimal worst-case contraction bound is

$$s(\zeta) = 1 - \frac{2}{1 + \frac{b^2}{a^2}}. \tag{88}$$

Our proof uses results on the structure of the Kronecker sum subspace to extend to our subspace restricted setting the methodology that Guillot et al. (2012) used to derive the unstructured GLasso convergence rates.

We decompose the claims of Theorem 25 into the following two theorems which we prove separately.

THEOREM 26. *Assume that the iterates Ω_t of Algorithm 4 satisfy $aI \preceq \Omega_t \preceq bI$, for all t, for some fixed constants $0 < a < b < \infty$. Suppose further that Ω^* is the global optimum. If $\zeta_t \leq a^2$ for all t, then*

$$\|\Omega_{t+1} - \Omega^*\|_F \leq \max\left\{\left|1 - \frac{\zeta_t}{b^2}\right|, \left|1 - \frac{\zeta_t}{a^2}\right|\right\} \|\Omega_t - \Omega^*\|_F.$$

Furthermore, the step size ζ_t which yields an optimal worst-case contraction bound $s(\zeta_t)$ is $\zeta = \frac{2}{a^{-2}+b^{-2}}$. The corresponding optimal worst-case contraction bound is

$$s(\zeta) = 1 - \frac{2}{1 + \frac{b^2}{a^2}}. \tag{89}$$

THEOREM 27. *Let $\rho_k \geq 0$ for all k and let Ω_{init} be the initialization of the TG-ISTA implementation of TeraLasso (Algorithm 4). Let*

$$a = \frac{1}{\sum_{k=1}^K \|S_k\|_2 + d_k \rho_k}, \qquad b = \|\Omega^*\|_2 + \|\Omega_{\mathrm{init}} - \Omega^*\|_F,$$

and assume $\zeta_t \leq a^2$ for all t. Then the iterates Ω_t of Algorithm 4 satisfy $aI \preceq \Omega_t \preceq bI$ for all t.

Observe that by Theorem 27, the worst case contraction factor (89)

$$s(\zeta) = 1 - \frac{2}{1 + (\|\Omega^*\|_2 + \|\Omega_{\mathrm{init}} - \Omega^*\|_F)^2 (\sum_{k=1}^K \|S_k\|_2 + d_k\rho_k)^2}$$

scales at most as $s(\zeta) = O(1 - \frac{2}{1+K^2})$ for $\|\Omega^*\|_2, \|\Sigma_0\|_2$ of fixed order, since $\|S_k\|_2 \sim \|\Sigma_0\|_2$ with high probability.

Let T be the number of iterations required for $\|\Omega_T - \Omega^*\|_F \leq \|\Omega^* - \widehat{\Omega}\|_F$ to hold, i.e. for the optimization error to be smaller than the statistical error. By Theorem 1, we require

$$\|\Omega_T - \Omega^*\|_F^2 \leq C_1 K^2 (s + p) \frac{\log p}{n \min_k m_k}. \tag{90}$$

Using worst case contraction factor $s(\zeta)$, (90) will hold for T such that (with high probability)

$$\|\Omega_{\mathrm{init}} - \Omega^*\|_F^2 \left(1 - \frac{2}{1 + \frac{b^2}{a^2}}\right)^{2T} \leq C_1 K^2 (s+p) \frac{\log p}{n \min_k m_k}.$$

Taking the logarithm of both sides and using $s(\zeta) = O(1 - \frac{2}{1+K^2})$, we have that the optimization error is guaranteed to equal the statistical error after T iterations, where

$$T = O_p\left(\frac{2\log K + \log(s+p) + \log\log p - \log(n \min_k m_k)}{\log\left(1 - \frac{2}{1+K^2}\right)}\right).$$

8.1. Proof of Theorem 26

For convenience, define the Kronecker sum shrinkage operator as

$$\text{shrink}_\rho^-(A) = \text{shrink}_{\rho_1}^-(A^{(1)}) \oplus \cdots \oplus \text{shrink}_{\rho_K}^-(A^{(K)}) \tag{91}$$

for $A = A^{(1)} \oplus \cdots \oplus A^{(K)} \in \mathcal{K}_\mathbf{p}$ and $\rho = [\rho_1, \dots, \rho_K]$ with all $\rho_k \geq 0$. Note that $\text{shrink}_\rho^-(A) = \arg\min_{\Omega \in \mathcal{K}_\mathbf{p}} \left\{ \frac{1}{2} \|\Omega - A\|_F^2 + \sum_{k=1}^K m_k \rho_k |\Psi_k|_{1,\text{off}} \right\}$. Since $\sum_{k=1}^K m_k \rho_k |\Psi_k|_{1,\text{off}}$ is a convex function on $\mathcal{K}_\mathbf{p}$, and since $\mathcal{K}_\mathbf{p}$ is a linear subspace, $\text{shrink}_\epsilon^-(\cdot)$ is a proximal operator by definition.

Recall that we can write the TG-ISTA update (28) using this Kronecker sum shrinkage operator as

$$\Omega_{t+1} = \arg\min_{\Omega \in \mathcal{K}_\mathbf{p}} \left\{ \frac{1}{2} \left\| \Omega - \left(\Omega_t - \zeta_t \left(\widetilde{S} - G^t \right) \right) \right\|_F^2 + \zeta_t \sum_{k=1}^K m_k \rho_k |\Psi_k|_{1,\text{off}} \right\}$$

$$= \arg\min_{\Omega \in \mathcal{K}_\mathbf{p}} \left\{ \frac{1}{2} \left\| \Omega - \left(\Omega_t - \zeta_t \left(\text{Proj}_{\mathcal{K}_\mathbf{p}}(\widehat{S} - \Omega_t^{-1} \right) \right) \right\|_F^2 + \zeta_t \sum_{k=1}^K m_k \rho_k |\Psi_k|_{1,\text{off}} \right\}$$

$$= \text{shrink}_{\zeta_t \rho}^-(\Omega_t - \zeta_t \text{Proj}_{\mathcal{K}_\mathbf{p}}(\widehat{S} - \Omega_t^{-1})),$$

where $\widehat{S}$ is the sample covariance (3) and $\widetilde{S} = \text{Proj}_{\mathcal{K}_\mathbf{p}}(\widehat{S})$ is its projection onto $\mathcal{K}_\mathbf{p}$ (15).

By convexity in $\mathcal{K}_\mathbf{p}$ and Theorem 6, the optimal point Ω_ρ^* is a fixed point of the ISTA iteration (Combettes and Wajs (2005), Prop 3.1). Thus,

$$\Omega_\rho^* = \text{shrink}_{\zeta_t \rho}^-(\Omega_\rho^* - \zeta_t \text{Proj}_{\mathcal{K}_\mathbf{p}}(\widehat{S} - (\Omega_\rho^*)^{-1}).$$

Since proximal operators are not expansive (Combettes and Wajs, 2005), we have

$$\begin{aligned}
&\|\Omega_{t+1} - \Omega_\rho^*\|_F \\
&\quad = \Big\| \text{shrink}_{\zeta_t \rho}^-(\Omega_t - \zeta_t \text{Proj}_{\mathcal{K}_\mathbf{p}}(\widehat{S} - \Omega_t^{-1})) \\
&\qquad\qquad - \text{shrink}_{\zeta_t \rho}^-(\Omega_\rho^* - \zeta_t \text{Proj}_{\mathcal{K}_\mathbf{p}}(\widehat{S} - (\Omega_\rho^*)^{-1})) \Big\|_F \\
&\quad \leq \|\Omega_t - \zeta_t \text{Proj}_{\mathcal{K}_\mathbf{p}}(\widehat{S} - \Omega_t^{-1}) - (\Omega_\rho^* - \zeta_t \text{Proj}_{\mathcal{K}_\mathbf{p}}(\widehat{S} - (\Omega_\rho^*)^{-1}))\|_F \\
&\quad = \|\Omega_t + \zeta_t \text{Proj}_{\mathcal{K}_\mathbf{p}}(\Omega_t^{-1}) - (\Omega_\rho^* + \zeta_t \text{Proj}_{\mathcal{K}_\mathbf{p}}((\Omega_\rho^*)^{-1}))\|_F.
\end{aligned}$$

For $\gamma > 0$ define $h_\gamma : \mathcal{K}_\mathbf{p}^\sharp \to \mathcal{K}_\mathbf{p}^\sharp$ by

$$h_\gamma(\Omega) = \text{vec}(\Omega) + \text{vec}(\gamma \text{Proj}_{\mathcal{K}_\mathbf{p}}(\Omega^{-1})).$$

Since $\partial \Omega^{-1} / \partial \Omega = -\Omega^{-1} \otimes \Omega^{-1}$,

$$\frac{\partial \text{Proj}_{\mathcal{K}_\mathbf{p}}(\Omega^{-1})}{\partial \Omega} = -P(\Omega^{-1} \otimes \Omega^{-1})P^T$$

where P is the projection matrix that projects $\mathrm{vec}(\Omega)$ onto the vectorized subspace $\mathcal{K}_{\mathbf{p}}$. Thus, we have the Jacobian (valid for all $\Omega \in \mathcal{K}_{\mathbf{p}}^{\sharp}$)

$$J_{h_\gamma}(\Omega) = PP^T - \gamma P(\Omega^{-1} \otimes \Omega^{-1})P^T.$$

Recall that if $h : U \subset \mathbb{R}^n \to \mathbb{R}^m$ is a differentiable mapping, then if $x, y \in U$ and U is convex, then if $J_h(\cdot)$ is the Jacobian of h,

$$\|h(x) - h(y)\| \leq \sup_{c\in[0,1]} \|J_h(cx + (1-c)y)\|\|x - y\|.$$

Thus, letting $Z_{t,c} = \mathrm{vec}(c\Omega_t + (1-c)\Omega^*_\rho)$, for $c \in [0,1]$ we have

$$\|h_{\zeta_t}(x) - h_{\zeta_t}(y)\| \leq \sup_{c\in[0,1]} \|PP^T - \zeta_t P(Z_{t,c}^{-1} \otimes Z_{t,c}^{-1})P^T\|\|\Omega_t - \Omega^*_\rho\|_F.$$

By Weyl's inequality, $\lambda_{\max}(Z_{t,c}) \leq \max\{\|\Omega_t\|, \|\Omega^*_\rho\|\}$ and

$$\lambda_{\min}(Z_{t,c}) \geq \min\{\lambda_{\min}(\Omega_t), \lambda_{\min}(\Omega^*_\rho)\}.$$

Furthermore, note that for any Y and projection matrix P

$$\lambda_{\max}(PYP^T) \leq \lambda_{\max}(Y).$$

We then have

$$\begin{aligned}|PP^T - \zeta_t P(Z_{t,c}^{-1} \otimes Z_{t,c}^{-1})P^T\| &\leq \|I_{p^2} - \zeta_t Z_{t,c}^{-1} \otimes Z_{t,c}^{-1}\| \\ &\leq \max\left\{\left|1 - \frac{\zeta_t}{b^2}\right|, \left|1 - \frac{\zeta_t}{a^2}\right|\right\},\end{aligned}$$

where the latter inequality comes from (Guillot et al., 2012). Thus,

$$\begin{aligned}\|\Omega_{t+1} - \Omega^*_\rho\|_F &\leq s(\zeta_t)\|\Omega_t - \Omega^*_\rho\|_F \\ \text{and} \quad s(\zeta) &= \max\left\{\left|1 - \frac{\zeta}{b^2}\right|, \left|1 - \frac{\zeta}{a^2}\right|\right\}\end{aligned}$$

as desired. Algorithm 4 will then converge if $s(\zeta_t) \in (0,1)$ for all t. The minimum of $s(\zeta)$ occurs at $\zeta = \frac{2}{a^{-2}+b^{-2}}$, completing the proof of Theorem 26. □

8.2. Proof of Theorem 27

We first prove the following properties of the Kronecker sum projection operator.

LEMMA 28. *For any $A \in \mathbb{R}^{p\times p}$ and orthogonal matrices $U_k \in \mathbb{R}^{d_k \times d_k}$, let $U = U_1 \otimes \cdots \otimes U_K \in \mathcal{K}_{\mathbf{p}}$. Then*

$$\mathrm{Proj}_{\mathcal{K}_{\mathbf{p}}}(A) = U\mathrm{Proj}_{\mathcal{K}_{\mathbf{p}}}(U^T AU)U^T.$$

Furthermore, if the eigendecomposition of A is of the form $A = (U_1 \otimes \cdots \otimes U_K)\Lambda(U_1 \otimes \cdots \otimes U_K)^T$ with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$, we have

$$\mathrm{Proj}_{\mathcal{K}_{\mathbf{p}}}(A) = U\mathrm{Proj}_{\mathcal{K}_{\mathbf{p}}}(\Lambda)U^T$$

and

$$\lambda_{\min}(A) \le \lambda_{\min}(\mathrm{Proj}_{\mathcal{K}_{\mathbf{p}}}(A)) \le \lambda_{\max}(\mathrm{Proj}_{\mathcal{K}_{\mathbf{p}}}(A)) \le \lambda_{\max}(A).$$

PROOF. Recall

$$\mathrm{Proj}_{\mathcal{K}_{\mathbf{p}}}(A) = \arg\min_{M\in\mathcal{K}_{\mathbf{p}}} \|A - B\|_F^2 = \arg\min_{B\in\mathcal{K}_{\mathbf{p}}} \|U^T AU - U^T BU\|_F^2$$

since $U^T AU = \Lambda$ and the Frobenius norm is unitarily invariant. Now, note that for any matrix $B = B_1 \oplus \cdots \oplus B_K \in \mathcal{K}_{\mathbf{p}}$,

$$\begin{aligned}
&(U_1 \otimes \ldots \otimes U_K)^T B(U_1 \otimes \cdots \otimes U_K)\\
&\quad= \sum_{k=1}^{K} (U_1 \otimes \cdots \otimes U_K)^T (I_{[d_{1:k-1}]} \otimes B_k \otimes I_{[d_{k+1}:K]})(U_1 \otimes \cdots \otimes U_K)\\
&\quad= \sum_{k=1}^{K} I_{[d_{1:k-1}]} \otimes U_k^T B_k U_k \otimes I_{[d_{k+1}:K]}\\
&\quad= (U_1^T B_1 U_1) \oplus \cdots \oplus (U_K^T B_K U_K)\\
&\quad\in \mathcal{K}_{\mathbf{p}},
\end{aligned}$$

since $U_k^T I_{d_k} U_k = I_{d_k}$. Since $U^T BU \in \mathcal{K}_{\mathbf{p}}$, the constraint $B \in \mathcal{K}_{\mathbf{p}}$ can be moved to $C = U^T BU$, giving

$$\begin{aligned}
\mathrm{Proj}_{\mathcal{K}_{\mathbf{p}}}(A) &= U(\arg\min_{C\in\mathcal{K}_{\mathbf{p}}} \|U^T AU - C\|_F^2)U^T\\
&= U(\mathrm{Proj}_{\mathcal{K}_{\mathbf{p}}}(U^T AU))U^T.
\end{aligned}$$

If $A = (U_1 \otimes \cdots \otimes U_K)\Lambda(U_1 \otimes \cdots \otimes U_K)^T$, then $U^T AU = \Lambda$, completing the first part of the proof. As shown in Lemma 33, $\mathrm{Proj}_{\mathcal{K}_{\mathbf{p}}}(\Lambda)$ is a diagonal matrix whose entries are weighted averages of the diagonal elements λ_i. Hence

$$\min_i \lambda_i \le \min_i [\mathrm{Proj}_{\mathcal{K}_{\mathbf{p}}}(\Lambda)]_{ii} \le \max_i [\mathrm{Proj}_{\mathcal{K}_{\mathbf{p}}}(\Lambda)]_{ii} \le \max_i \lambda_i.$$

Since $\mathrm{Proj}_{\mathcal{K}_{\mathbf{p}}}(\Lambda)$ gives the eigenvalues of $\mathrm{Proj}_{\mathcal{K}_{\mathbf{p}}}(A)$ by the orthogonality of U, this completes the proof. □

LEMMA 29. *Let* $0 < a < b$ *be given positive constants and let* $\zeta_t > 0$. *Assume* $aI \preceq \Omega_t \preceq bI$. *Then for*

$$\Omega_{t+1/2} := \Omega_t - \zeta_t(\mathrm{Proj}_{\mathcal{K}_\mathbf{p}}(\widehat{S} - \Omega_t^{-1}))$$

we have

$$\lambda_{\min}(\Omega_{t+1/2}) \geq \begin{cases} 2\sqrt{\zeta_t} - \zeta_t\lambda_{\max}(\widehat{S}) & \text{if } a \leq \sqrt{\zeta_t} \leq b \\ \min\left(a + \frac{\zeta_t}{a}, b + \frac{\zeta_t}{b}\right) - \zeta_t\lambda_{\max}(\widehat{S}) & \text{o.w.} \end{cases}$$

PROOF. Let $U\Gamma U^T = \Omega_t$ be the eigendecomposition of Ω_t, where $\Gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_p)$. Then all $b \geq \gamma_i \geq a > 0$. Since $\Omega_t \in \mathcal{K}_\mathbf{p}$, by the eigendecomposition property in Appendix A we have $U = U_1 \otimes \cdots \otimes U_K$ and $\Gamma \in \mathcal{K}_\mathbf{p}$, letting us apply Lemma 28:

$$\begin{aligned} \Omega_{t+1/2} &= \Omega_t - \zeta_t(\mathrm{Proj}_{\mathcal{K}_\mathbf{p}}(\widehat{S}) - \mathrm{Proj}_{\mathcal{K}_\mathbf{p}}(\Omega_t^{-1})) \\ &= U\Gamma U^T - \zeta_t(\mathrm{Proj}_{\mathcal{K}_\mathbf{p}}(\widehat{S}) - U\mathrm{Proj}_{\mathcal{K}_\mathbf{p}}(\Gamma^{-1})U^T) \\ &= U\left(\Gamma - \zeta_t(U^T\mathrm{Proj}_{\mathcal{K}_\mathbf{p}}(\widehat{S})U - \mathrm{Proj}_{\mathcal{K}_\mathbf{p}}(\Gamma^{-1}))\right)U^T \\ &= U\left(\mathrm{Proj}_{\mathcal{K}_\mathbf{p}}(\Gamma) - \zeta_t\left(\mathrm{Proj}_{\mathcal{K}_\mathbf{p}}(U^T\widehat{S}U) - \mathrm{Proj}_{\mathcal{K}_\mathbf{p}}(\Gamma^{-1})\right)\right)U^T \\ &= \mathrm{Proj}_{\mathcal{K}_\mathbf{p}}\left(U(\Gamma + \zeta\Gamma^{-1} - \zeta_t(U^T\widehat{S}U))U^T\right) \\ &= \mathrm{Proj}_{\mathcal{K}_\mathbf{p}}(\widetilde{\Omega}_{t+1/2}), \end{aligned}$$

where we set $\widetilde{\Omega}_{t+1/2} = U(\Gamma + \zeta\Gamma^{-1} - \zeta_t(U^T\widehat{S}U))U^T$ and recall the linearity of the projection operator $\mathrm{Proj}_{\mathcal{K}_\mathbf{p}}(\cdot)$ (Lemma 33). By Weyl's inequality,

$$\gamma_1 + \frac{\zeta_t}{\gamma_1} - \zeta_t\lambda_{\max}(\widehat{S}) \leq \lambda_{\min}(\widetilde{\Omega}_{t+1/2}).$$

By Lemma 28,

$$\gamma_1 + \frac{\zeta_t}{\gamma_1} - \zeta_t\lambda_{\max}(\widehat{S}) \leq \lambda_{\min}(\Omega_{t+1/2}).$$

Note that the only extremum of the function $f(x) = x + \frac{\zeta_t}{x}$ over $a \leq x \leq b$ is a global minimum at $x = \sqrt{\zeta_t}$. Hence

$$\inf_{a \leq x \leq b} x + \frac{\zeta_t}{x} = \begin{cases} 2\sqrt{\zeta_t} & \text{if } a \leq \sqrt{\zeta_t} \leq b \\ \min\left(a + \frac{\zeta_t}{a}, b + \frac{\zeta_t}{b}\right) & \text{o.w.} \end{cases}$$

By our assumption, $a \leq \gamma_1 \leq b$. Thus

$$\lambda_{\min}(\Omega_{t+1/2}) \geq \begin{cases} 2\sqrt{\zeta_t} - \zeta_t\lambda_{\max}(\widehat{S}) & \text{if } a \leq \sqrt{\zeta_t} \leq b \\ \min\left(a + \frac{\zeta_t}{a}, b + \frac{\zeta_t}{b}\right) - \zeta_t\lambda_{\max}(\widehat{S}) & \text{o.w.} \end{cases}$$

as desired, completing the proof. □

We then have the following lemma.

LEMMA 30. *For $A \in \mathcal{K}_{\mathbf{p}}^{\sharp}$ and $\epsilon = [\epsilon_1, \dots, \epsilon_K]$ with $\epsilon_k \geq 0$:*

$$\lambda_{\min}(A) - \sum_{k=1}^{K} d_k \epsilon_k \leq \lambda_{\min}(\mathrm{shrink}_{\epsilon}^{-}(A))$$

PROOF. Since by definition (91)

$$\mathrm{shrink}_{\epsilon}^{-}(A) = \mathrm{shrink}_{\epsilon_1}^{-}(A^{(1)}) \oplus \cdots \oplus \mathrm{shrink}_{\epsilon_K}^{-}(A^{(K)}),$$

we can use the fact that the eigenvalues of a Kronecker sum are the sums of the eigenvalues to show

$$\lambda_{\min}(\mathrm{shrink}_{\epsilon}^{-}(A)) = \sum_{k=1}^{K} \lambda_{\min}(\mathrm{shrink}_{\epsilon_k}^{-}(A^{(k)})).$$

We have used the fact that A is positive definite since it is in $\mathcal{K}_{\mathbf{p}}^{\sharp}$.

Via Weyl's inequality and the proof of Lemma 6 in (Guillot et al., 2012),

$$\lambda_{\min}(\mathrm{shrink}_{\epsilon_k}^{-}(A^{(k)})) \geq \lambda_{\min}(A^{(k)}) - d_k \epsilon_k.$$

Hence,

$$\lambda_{\min}(\mathrm{shrink}_{\epsilon}^{-}(A)) \geq \sum_{k=1}^{K} \lambda_{\min}(A^{(k)}) - \sum_{k=1}^{K} d_k \epsilon_k = \lambda_{\min}(A) - \sum_{k=1}^{K} d_k \epsilon_k$$

□

8.2.1. Proof of Theorem 27

To prove the lower inequality in Theorem 27, we show the following.

LEMMA 31. *Let $\rho = [\rho_1, \dots, \rho_K]$ with all $\rho_i > 0$. Define*

$$\chi = \sum_{k=1}^{K} d_k \rho_k$$

and let $\alpha = \frac{1}{\|\widehat{S}\|_2 + \chi} < b'$. Assume $\alpha I \preceq \Omega_{t+1}$. Then $\alpha I \preceq \Omega_{t+1}$ for every $0 < \zeta_t < \alpha^2$.

PROOF. Since $\zeta_t < \alpha^2$, $\sqrt{\zeta_t} \notin [\alpha, b']$, and $\min\left(\alpha + \frac{\zeta_t}{\alpha}, b' + \frac{\zeta_t}{b'}\right) = \alpha + \frac{\zeta_t}{\alpha}$. Lemma 29 then implies that

$$\begin{aligned}\lambda_{\min}(\Omega_{t+1/2}) &\geq \min\left(\alpha + \frac{\zeta_t}{\alpha}, b' + \frac{\zeta_t}{b'}\right) - \zeta_t \lambda_{\max}(\widehat{S}) \\ &= \alpha + \frac{\zeta_t}{\alpha} - \zeta_t \lambda_{\max}(\widehat{S}).\end{aligned}$$

By Lemma 30,

$$\begin{aligned}\lambda_{\min}(\Omega_{t+1}) &= \lambda_{\min}\left(\text{shrink}^-_{\zeta_t\rho}(\Omega_{t+1/2})\right)\\ &\geq \lambda_{\min}(\Omega_{t+1/2}) - \zeta_t\chi\\ &\geq \alpha + \frac{\zeta_t}{\alpha} - \zeta_t\lambda_{\max}(\widehat{S}) - \zeta_t\chi.\end{aligned}$$

Hence, since $\zeta_t > 0$, $\lambda_{\min}(\Omega_{t+1}) \geq \alpha$ whenever

$$\begin{aligned}\zeta_t\left(\frac{1}{\alpha} - \lambda_{\max}(\widehat{S}) - \chi\right) &\geq 0\\ \frac{1}{\alpha} - \lambda_{\max}(\widehat{S}) - \chi &\geq 0\\ \alpha &\leq \frac{1}{\|\widehat{S}\|_2 + \chi}.\end{aligned}$$

□

The upper bound in Theorem 27 results from the following lemma.

LEMMA 32. *Let χ be as in Lemma 31 and let $\alpha = \frac{1}{\|\widehat{S}\|_2+\chi}$. Let $\zeta_t \leq \alpha^2$ for all t. We then have $\Omega_t \preceq b'I$ for all t when $b' = \|\Omega^*_\rho\|_2 + \|\Omega_0 - \Omega^*_\rho\|_F$.*

PROOF. By Lemma 31, $\alpha I \preceq \Omega_t$ for every t. Since $\Omega_t \to \Omega^*_\rho$, by strong convexity $\alpha I \preceq \Omega^*_\rho$. Hence $a = \min\{\lambda_{\min}(\Omega_t), \lambda_{\min}(\Omega^*_\rho)\} \geq \alpha$. For $b > a$ and $\zeta_t \leq \alpha^2$,

$$\max\left\{\left|1 - \frac{\zeta_t}{b^2}\right|, \left|1 - \frac{\zeta_t}{a^2}\right|\right\} \leq 1.$$

Hence, by Theorem 25 $\|\Omega_t - \Omega^*_\rho\|_F \leq \|\Omega_{t-1} - \Omega^*_\rho\|_F \leq \|\Omega_0 - \Omega^*_\rho\|_F$. Thus

$$\|\Omega_t\|_2 - \|\Omega^*_\rho\|_2 \leq \|\Omega_t - \Omega^*_\rho\|_2 \leq \|\Omega_t - \Omega^*_\rho\|_F \leq \|\Omega_0 - \Omega^*_\rho\|_F$$

so

$$\|\Omega_t\|_2 \leq \|\Omega^*_\rho\|_2 + \|\Omega_0 - \Omega^*_\rho\|_F.$$

□

This completes the proof of Theorem 27. □

A. Useful Properties of the Kronecker Sum and $\mathcal{K}_{\mathbf{p}}$

A.1. Basic Properties

As the properties of Kronecker sums are not always widely known, we have compiled a list of some fundamental algebraic relations we use.

(a) Sum or difference of Kronecker sums (Laub, 2005):

$$\begin{aligned} c_A(A_1 \oplus \cdots \oplus A_K) + c_B(B_1 \oplus \cdots \oplus B_K) \\ = (c_A A_1 + c_B B_1) \oplus \cdots \oplus (c_A A_K + c_B B_K). \end{aligned}$$

(b) Factor-wise disjoint off diagonal support (Laub, 2005). By construction, if for any k and $i \neq j$

$$[I_{[d_{1:k-1}]} \otimes A_k \otimes I_{[d_{k+1:K}]}]_{ij} \neq 0,$$

then for all $\ell \neq k$

$$[I_{[d_{1:\ell-1}]} \otimes A_\ell \otimes I_{[d_{\ell+1:K}]}]_{ij} = 0.$$

Thus,

$$|A_1 \oplus \cdots \oplus A_K|_{1,\text{off}} = \sum_{k=1}^{K} |I_{[d_{1:k-1}]} \otimes \text{offd}(A_k) \otimes I_{[d_{k+1:K}]}|_1 = \sum_{k=1}^{K} m_k |A_k|_{1,\text{off}}.$$

(c) Eigendecomposition: If $A_k = U_k \Lambda_k U_k^T$ are the eigendecompositions of the factors, then (Laub, 2005)

$$A_1 \oplus \cdots \oplus A_K = (U_1 \otimes \cdots \otimes U_K)(\Lambda_1 \oplus \cdots \oplus \Lambda_K)(U_1 \otimes \cdots \otimes U_K)^T$$

is the eigendecomposition of $A_1 \oplus \cdots \oplus A_K$. Some resulting identities useful for doing numerical calculations are as follows:

(i) L2 norm:

$$\begin{aligned} \|A_1 \oplus \cdots \oplus A_K\|_2 &= \max\left(\sum_{k=1}^{K} \max_i [\Lambda_k]_{ii}, -\sum_{k=1}^{K} \min_i [\Lambda_k]_{ii}\right) \\ &\leq \sum_{k=1}^{K} \|A_k\|_2. \end{aligned}$$

(ii) Determinant:

$$\begin{aligned} \log|A_1 \oplus \cdots \oplus A_K| &= \log|\Lambda_1 \oplus \cdots \oplus \Lambda_K| \\ &= \underbrace{\sum_{i_1=1}^{d_1} \cdots \sum_{i_K=1}^{d_K}}_{\text{K sums}} \log\left(\underbrace{[\Lambda_1]_{i_1 i_1} + \cdots + [\Lambda_K]_{i_K i_K}}_{\text{K terms}}\right). \end{aligned}$$

(iii) Matrix powers (e.g. inverse, inverse square root):

$$(A_1 \oplus \cdots \oplus A_K)^v = (U_1 \otimes \cdots \otimes U_K)(\Lambda_1 \oplus \cdots \oplus \Lambda_K)^v (U_1 \otimes \cdots \otimes U_K)^T.$$

Since the Λ_k are diagonal, this calculation is memory and computation efficient.

A.2. Eigenstructure of $\Omega \in \mathcal{K}_{\mathbf{p}}$

Kronecker sum matrices $\Omega \in \mathcal{K}_{\mathbf{p}}$ have Kronecker product eigenvectors with linearly related eigenvalues, as contrasted to the multiplicatively related eigenvalues in the Kronecker product. For simplicity, we illustrate in the $K = 2$ case, but the result generalizes to the full tensor case. Suppose that $\Psi_1 = U_1 \Lambda_1 U_1^T$ and $\Psi_2 = U_2 \Lambda_2 U_2^T$ are the eigendecompositions of Ψ_1 and Ψ_2. Then by Laub (2005), if $\Omega = \Psi_1 \oplus \Psi_2$, the eigendecomposition of Ω is

$$\Omega = \Psi_1 \oplus \Psi_2 = (U_1 \otimes U_2)(\Lambda_1 \oplus \Lambda_2)(U_1 \otimes U_2)^T.$$

Thus, the eigenvectors of the Kronecker sum are the Kronecker products of the eigenvectors of each factor. This "block" structure is evident in the inverse Kronecker sum example in Section 1 of the main text. The structure of Ω^{-1} is discussed further in Canuto et al. (2014).

This eigenstructure representation parallels the eigenvector structure of the Kronecker product - specifically when $\Omega = \Psi_1 \otimes \Psi_2$

$$\Omega = \Psi_1 \otimes \Psi_2 = (U_1 \otimes U_2)(\Lambda_1 \otimes \Lambda_2)(U_1 \otimes U_2)^T.$$

Hence, use of the Kronecker sum model can be viewed as replacing the nonconvex, multiplicative eigenvalue structure of the Kronecker product with the convex linear eigenvalue structure of the Kronecker sum. This additive structure results in relatively more stable estimation of the precision matrix. As the tensor dimension K increases, this structural stability of the Kronecker sum as compared to the Kronecker product becomes more dominant (K term sums instead of K-order products).

A.3. Projection onto $\mathcal{K}_{\mathbf{p}}$

We first introduce a submatrix notation. Fix a k, and choose $i, j \in \{1, \dots m_k\}$. Let $E_1 \in \mathbb{R}^{\prod_{\ell=1}^{k-1} d_k \times \prod_{\ell=1}^{k-1} d_k}$ and $E_2 \in \mathbb{R}^{\prod_{\ell=k+1}^{K} d_k \times \prod_{\ell=k+1}^{K} d_k}$ be such that $[E_1 \otimes E_2]_{ij} = 1$ with all other elements zero. Observe that $E_1 \otimes E_2 \in \mathbb{R}^{m_k \times m_k}$. For any matrix $A \in \mathbb{R}^{p \times p}$, let $A(i, j|k) \in \mathbb{R}^{d_k \times d_k}$ be the submatrix of A defined via

$$[A(i,j|k)]_{rs} = \mathrm{tr}((E_1 \otimes \mathbf{e}_r \mathbf{e}_s \otimes E_2)A), \qquad r, s = 1, \dots, d_k. \tag{92}$$

The submatrix $A(i, j|k)$ is defined for all $i, j \in \{1, \dots m_k\}$ and $k = 1, \dots, K$. When A is a covariance matrix associated with a tensor X, this subblock corresponds to the covariance matrix between the ith and jth slices of X along the kth dimension.

We can now express the projection operator $\mathrm{Proj}_{\mathcal{K}_{\mathbf{p}}}(A)$ in closed form:

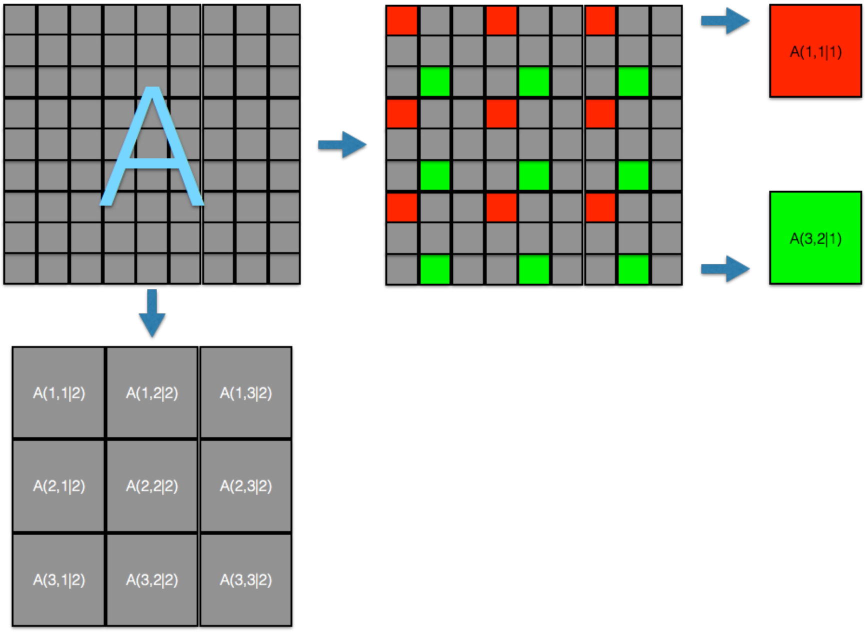


Fig. 19: Submatrix notation (equation (92)). Shown is a 9x9 matrix A, with $K = 2$ and $d_1 = d_2 = 3$. Displayed are the subblocks corresponding to the $A(i, j|2)$ and two example $A(i, j|1)$. $A(1, 1|1) \in \mathbb{R}^{3\times3}$ is formed from the 9 red entries, and $A(3, 2|1)$ from the nine green entries. The remaining $A(i, j|1)$ follow similarly according to (92).

LEMMA 33 (PROJECTION ONTO $\mathcal{K}_\mathbf{p}$). *For any $A \in \mathbb{R}^{p\times p}$,*

$$\begin{aligned}\mathrm{Proj}_{\mathcal{K}_\mathbf{p}}(A) &= A_1 \oplus \cdots \oplus A_K - (K-1)\frac{\mathrm{tr}(A)}{p} I_p \\ &= \left(A_1 - \frac{K-1}{K}\frac{\mathrm{tr}(A_1)}{d_1} I_{d_1}\right) \oplus \cdots \oplus \left(A_K - \frac{K-1}{K}\frac{\mathrm{tr}(A_K)}{d_K} I_{d_K}\right),\end{aligned}$$

where

$$A_k = \frac{1}{m_k}\sum_{i=1}^{m_k} A(i, i|k).$$

Since the submatrix operator $A(i, i|k)$ is clearly linear, $\mathrm{Proj}_{\mathcal{K}_\mathbf{p}}(\cdot)$ *is a linear operator.*

PROOF. Since $\mathcal{K}_\mathbf{p}$ is a linear subspace, projection can be found via inner products. Specifically, recall that if a subspace $\mathcal{A}$ is spanned by an orthonormal basis U, then

$$\mathrm{Proj}_{\mathcal{A}}(\mathbf{x}) = UU^T\mathbf{x}.$$

Since $\mathcal{K}_{\mathbf{p}}$ is the space of Kronecker sums, the off diagonal elements are independent and do not overlap across factors. The diagonal portion is more difficult as each factor overlaps on the same entries, creating an overdetermined system. We can create an alternate parameterization of $\mathcal{K}_{\mathbf{p}}$:

$$\mathrm{Proj}_{\mathcal{K}_{\mathbf{p}}}(A) = \bar{A}_1 \oplus \cdots \oplus \bar{A}_K + \tau_A I_p = \tau_A I_p + \sum_{k=1}^{K} I_{[d_{1:k-1}]} \otimes \bar{A}_k \otimes I_{[d_{k+1:K}]} \tag{93}$$

where we constrain $\mathrm{tr}(\bar{A}_k) = 0$. Each of the $K+1$ terms in this sum is now orthogonal to all other terms since by construction

$$\begin{aligned}
&\langle I_{[d_{1:k-1}]} \otimes \bar{A}_k \otimes I_{[d_{k+1:K}]}, I_{[d_{1:\ell-1}]} \otimes \bar{A}_\ell \otimes I_{[d_{\ell+1:K}]} \rangle \\
&\qquad = \frac{p}{d_k d_\ell} \mathrm{tr}((\bar{A}_k \otimes I_{d_\ell})(I_{d_k} \otimes \bar{A}_\ell)) = \frac{p}{d_k d_\ell} \mathrm{tr}(\bar{A}_k)\mathrm{tr}(\bar{A}_\ell) = 0 \\
&\langle \tau_A I_p, I_{[d_{1:k-1}]} \otimes \bar{A}_k \otimes I_{[d_{k+1:K}]} \rangle \\
&\qquad = \langle \tau_A I_{[d_{1:k-1}]} \otimes I_{d_k} \otimes I_{[d_{k+1:K}]}, I_{[d_{1:k-1}]} \otimes \bar{A}_k \otimes I_{[d_{k+1:K}]} \rangle \\
&\qquad = m_k \langle I_{d_k}, \bar{A}_k \rangle = m_k \mathrm{tr}(\bar{A}_k) = 0
\end{aligned}$$

for $\ell \neq k$ and all possible $\bar{A}_k$, τ_A. Thus, we can form bases for the $\bar{A}_k$ and τ_A independently. To find the $\bar{A}_k$ it suffices to project A onto a basis for $\bar{A}_k$. We can divide this projection into two steps. In the first step, we ignore the constraint on $\mathrm{tr}(\bar{A}_k)$ and create the orthonormal basis

$$\mathbf{u}_k^{(ij)} := \frac{1}{\sqrt{m_k}} I_{[d_{1:k-1}]} \otimes \mathbf{e}_i \mathbf{e}_j^T \otimes I_{[d_{k+1:K}]}$$

for all $i, j = 1, \ldots d_k$. Recall that in a projection of $\mathbf{x}$, the coefficient of a basis component $\mathbf{u}$ is given by $\mathbf{u}^T \mathbf{x} = \langle \mathbf{u}, \mathbf{x} \rangle$. We can thus apply this elementwise to the projection of A. Hence projecting A onto these basis components yields a matrix $B\sqrt{m_k} \in \mathbb{R}^{d_k \times d_k}$ where

$$B_{ij} = \frac{1}{m_k} \langle A, I_{[d_{1:k-1}]} \otimes \mathbf{e}_i \mathbf{e}_j^T \otimes I_{[d_{k+1:K}]} \rangle.$$

To enforce the $\mathrm{tr}(\bar{A}_k) = 0$ constraint, we project away from B the one-dimensional subspace spanned by I_{d_k}. This projection is given by

$$B - \frac{\mathrm{tr}(B)}{d_k} I_{d_k}, \tag{94}$$

where by construction

$$\begin{aligned}
\frac{\mathrm{tr}(B)}{d_k} &= \frac{1}{d_k m_k} \sum_{i=1}^{d_k} \langle A, I_{[d_{1:k-1}]} \otimes \mathbf{e}_i \mathbf{e}_i^T \otimes I_{[d_{k+1:K}]} \rangle \\
&= \frac{1}{p} \langle A, I_p \rangle = \frac{\mathrm{tr}(A)}{p}.
\end{aligned}$$

Equation (94) completes the projection onto a basis for $\bar{A}_k$, so we can expand the projection $\sqrt{m_k}B$ back into the original space. This yields a $\bar{A}_k$ of the form

$$[\bar{A}_k]_{ij} = \begin{cases} \frac{1}{m_k}\langle A, I_{[d_{1:k-1}]} \otimes \mathbf{e}_i\mathbf{e}_j^T \otimes I_{[d_{k+1:K}]}\rangle & i \neq j \\ \frac{1}{m_k}\langle A, I_{[d_{1:k-1}]} \otimes \mathbf{e}_i\mathbf{e}_i^T \otimes I_{[d_{k+1:K}]}\rangle - \frac{\mathrm{tr}(A)}{p} & i = j \end{cases}$$

Finally, for τ_A we can compute

$$\tau_A = \frac{1}{p}\langle A, I_p\rangle = \frac{\mathrm{tr}(A)}{p}.$$

Combining all these together and substituting into (93) allows us to define the projection in terms of matrices $\widetilde{A}_k$, where we split the $\tau_A I_p$ term evenly across the other K factors. Specifically

$$\mathrm{Proj}_{\mathcal{K}_\mathbf{p}}(A) = \widetilde{A}_1 \oplus \cdots \oplus \widetilde{A}_K.$$

where

$$[\widetilde{A}_k]_{ij} = \begin{cases} \frac{1}{m_k}\langle A, I_{[d_{1:k-1}]} \otimes \mathbf{e}_i\mathbf{e}_j^T \otimes I_{[d_{k+1:K}]}\rangle & i \neq j \\ \frac{1}{m_k}\langle A, I_{[d_{1:k-1}]} \otimes \mathbf{e}_i\mathbf{e}_i^T \otimes I_{[d_{k+1:K}]}\rangle - \frac{K-1}{K}\frac{\mathrm{tr}(A)}{p} & i = j \end{cases}. \tag{95}$$

An equivalent representation using factorwise averages is

$$\widetilde{A} = A_k - \frac{K-1}{K}\frac{\mathrm{tr}(A)}{p},$$

where

$$A_k = \frac{1}{m_k}\sum_{i=1}^{m_k} A(i, i|k).$$

Moving the trace corrections to a last term and putting the result in terms of the A_k yields the lemma.

In Algorithm 4 we use an efficient method of computing this projected inverse in our setting by exploiting the eigendecomposition identities in Section A.2. □

B. Known diagonal elements (correlation matrix form)

In the case where the diagonal $\mathrm{diag}(\Omega_0)$ of the precision matrix is known a priori, the estimation problem becomes easier. For simplicity, we consider the case that Ω_0 is in the form of a correlation matrix, i.e. $\mathrm{diag}(\Omega_0) = I_p$, noting this was the setting originally the focus of Kalaitzis et al. (2013).

Note that since the diagonal elements are known, we do not need to estimate them and indeed can set all the $\mathrm{diag}(\Psi_k) = 1/KI_{d_k}$. Revisiting the proof of Theorem 1, it is easy to show the following corollary, which shows strong $O(\sqrt{(K+1)s\frac{\log p}{n \min_k m_k}})$ convergence in the case of $\ell 1$ regularization. This replacement of the $\sqrt{p+s}$ term in rate of Theorem 1 with a $\sqrt{s}$ guarantees single sample convergence in the sparse setting when $\min_k m_k \gg s$.

COROLLARY 1. *Suppose the conditions of Theorem 1, and that* $\operatorname{diag}(\Omega_0) = I_p$ *is known. Then under event* $\mathcal{A}$,

$$\|\widehat{\Omega} - \Omega_0\|_F \leq \frac{2C_1 \|\Sigma_0\|_2}{\phi^2_{\min}(\Sigma_0)} \sqrt{(K+1)s\frac{\log p}{n \min_k m_k}}.$$

Furthermore, event $\mathcal{A}$ *holds with probability at least* $1 - 2(K+1)\exp(-c \log p)$.

PROOF. Dropping the diagonal term from the proof of Lemma 17, we have that the $\sqrt{p}$ dependence vanishes, and on event $\mathcal{A}$, we have $G(\Delta) > 0$ for all $\Delta \in \mathcal{T}_n$ where

$$\mathcal{T}_n = \{\Delta_\Omega \in \mathcal{K}_{\mathbf{p}} : \Delta_\Omega = \Omega - \Omega_0, \|\Delta_\Omega\|_F = M r_{n,\mathbf{p}}\}$$

and

$$r_{n,\mathbf{p}} = C \|\Sigma_0\|_2 \sqrt{s(K+1)\frac{\log p}{n \min_k m_k}}.$$

The rest of the proof follows by substituting this new value of $r_{n,\mathbf{p}}$ into the proof of Theorem 1.

C. SCAD and MCP regularizers

The SCAD penalty (Fan and Li, 2001) with parameter $a > 2$ (giving $\mu = 1/(a-1)$) is given by

$$g_\rho(t) = \begin{cases} \rho|t| & \text{if } |t| \leq \rho \\ -\frac{t^2 - 2a\rho|t| + \rho^2}{2(a-1)} & \text{if } \rho < |t| \leq a\rho \\ \frac{(a+1)\rho^2}{2} & \text{if } a\rho < |t| \end{cases} \tag{96}$$

which is linear (as the $\ell 1$ norm) for small $|t|$, constant for large $|t|$, and has a transition between the two regimes for moderate $|t|$.

The MCP penalty (Zhang et al., 2010) with parameter $a > 0$ (giving $\mu = 1/a$) is given by

$$g_\rho(t) = \operatorname{sign}(t)\rho \int_0^{|t|} \left(1 - \frac{z}{\rho a}\right)_+ dz, \tag{97}$$

giving a more smooth transition between the approximately linear region and the constant region ($t > \rho a$).

References

Barzilai, J. and Borwein, J. M. (1988) Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, **8**, 141–148.

Beck, A. and Teboulle, M. (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, **2**, 183–202.

Boyd, S. and Vandenberghe, L. (2009) *Convex optimization*. Cambridge university press.

Canuto, C., Simoncini, V. and Verani, M. (2014) On the decay of the inverse of matrices that are sum of kronecker products. *Linear Algebra and its Applications*, **452**, 21–39.

Combettes, P. L. and Wajs, V. R. (2005) Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, **4**, 1168–1200.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, **96**, 1348–1360.

Guillot, D., Rajaratnam, B., Rolfs, B., Maleki, A. and Wong, I. (2012) Iterative thresholding algorithm for sparse inverse covariance estimation. In *NIPS*, 1574–1582.

Hsieh, C.-J., Sustik, M. A., Dhillon, I. S. and Ravikumar, P. D. (2014) QUIC: quadratic approximation for sparse inverse covariance estimation. *Journal of Machine Learning Research*, **15**, 2911–2947.

Kalaitzis, A., Lafferty, J., Lawrence, N. and Zhou, S. (2013) The bigraphical lasso. In *Proceedings of the International Conference on Machine Learning*, 1229–1237.

Laub, A. J. (2005) *Matrix Analysis for Scientists and Engineers*. SIAM.

Loh, P.-L., Wainwright, M. J. et al. (2017) Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics*, **45**, 2455–2482.

Ortega, J. M. and Rheinboldt, W. C. (1970) *Iterative solution of nonlinear equations in several variables*, vol. 30. Siam.

Rothman, A. J., Bickel, P. J., Levina, E., Zhu, J. et al. (2008) Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, **2**, 494–515.

Rudelson, M., Vershynin, R. et al. (2013) Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, **18**.

Van Loan, C. and Pitsianis, N. (1993) Approximation with kronecker products. In *Linear Algebra for Large Scale and Real Time Applications*, 293–314. Kluwer Publications.

Zhang, C.-H. et al. (2010) Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, **38**, 894–942.

Zhou, S. (2014) Gemini: Graph estimation with matrix variate normal instances. *The Annals of Statistics*, **42**, 532–562.

Zhou, S., Lafferty, J. and Wasserman, L. (2010) Time varying undirected graphs. *Machine Learning*, **80**, 295–319.

Zhou, S., Rütimann, P., Xu, M. and Bühlmann, P. (2011) High-dimensional covariance estimation based on gaussian graphical models. *The Journal of Machine Learning Research*, **12**, 2975–3026.