S1 Appendix: Backward mapping for M-Estimator

The graphical model MLE can be expressed as a backward mapping \( H \) in an exponential family distribution that computes the model parameters corresponding to some given (sample) moments. There are however two caveats with this backward mapping: it is not available in closed form for many classes of models, and even if it were available in closed form, it need not be well-defined in high-dimensional settings (i.e., could lead to unbounded model parameter estimates).

We provide detailed explanations about backward mapping from the M-estimator framework \([2]\) and backward mapping for Gaussian special case in this section.

**Backward mapping:** Suppose a random variable \( X \in \mathbb{R}^p \) follows the exponential family distribution:

\[
\mathbb{P}(X; \theta) = h(X) \exp \{ \theta^T \phi(X) - A(\theta) \} \tag{S1–1}
\]

Where \( \theta \in \Theta \subseteq \mathbb{R}^d \) is the canonical parameter to be estimated and \( \Theta \) denotes the parameter space, \( \phi(X) \) denotes the sufficient statistics with a feature mapping function \( \phi: \mathbb{R}^p \rightarrow \mathbb{R}^d \), and \( A(\theta) \) is the log-partition function. We define mean parameters as: \( \nu(\theta) := \mathbb{E}[\phi(X)] \), which are the first moments of the sufficient statistics \( \phi(\theta) \) under the exponential family distribution. The set of all possible moments by the moment polytope:

\[
\mathcal{M} = \{ \nu | \exists p \text{ a distribution s.t. } \mathbb{E}_p[\phi(X)] = \nu \} \tag{S1–2}
\]

Most machine learning problems about graphical model inference involve the task of computing moments \( \nu(\theta) \in \mathcal{M} \) given the canonical parameters \( \theta \in \mathbb{H} \). We denote this computing as **forward mapping**:

\[
\mathcal{A}: \mathbb{H} \rightarrow \mathcal{M} \tag{S1–3}
\]

When we need to consider the reverse computing of the forward mapping, we denote the interior of \( \mathcal{M} \) as \( \mathcal{M}^0 \). The so-called **backward mapping** is defined as:

\[
\mathcal{A}^*: \mathcal{M}^0 \rightarrow \mathbb{H} \tag{S1–4}
\]

which does not need to be unique. For the exponential family distribution,

\[
\mathcal{A}^*: \nu(\theta) \rightarrow \theta = \nabla \mathcal{A}^*(\nu(\theta)). \tag{S1–5}
\]

Where \( \mathcal{A}^*(\nu(\theta)) = \sup_{\theta \in \mathbb{H}} \theta, \nu(\theta) > -\mathcal{A}(\theta) \).

**Backward Mapping: Gaussian Case** If the random variable \( X \in \mathbb{R}^p \) follows the Gaussian Distribution \( \mathcal{N}(\mu, \Sigma) \). Then \( \theta = (\Sigma^{-1} \mu, -\frac{1}{2} \Sigma^{-1}) \). The sufficient statistics \( \phi(X) = (X, XX^T) \) and the log-partition function \( A(\theta) = \frac{1}{2} \mu^T \Sigma^{-1} \mu + \frac{1}{2} \log(|\Sigma|) \). \( h(x) = (2\pi)^{-\frac{1}{2}} \).

When inferring the Gaussian Graphical Models, it is easy to estimate the mean vector \( \nu(\theta) \), since it equals to \( \mathbb{E}[X, XX^T] \).

Because the \( \theta \) contains entry \( \Sigma^{-1} \), when estimating sGGM, we need to use the backward mapping:

For the case of Gaussian distribution,

\[
\theta = (\Sigma^{-1} \mu, -\frac{1}{2} \Sigma^{-1}) = \mathcal{A}^*(\nu) = \nabla \mathcal{A}^*(\nu) = ((\mathbb{E}_\theta[X, XX^T] - \mathbb{E}_\theta[X] \mathbb{E}_\theta[X]^T)^{-1} \mathbb{E}_\theta[X], -\frac{1}{2} \mathbb{E}_\theta[X, XX^T] - \mathbb{E}_\theta[X] \mathbb{E}_\theta[X]^T)^{-1}). \tag{S1–6}
\]

By plugging in \( A(\theta) = \frac{1}{2} \mu^T \Sigma^{-1} \mu + \frac{1}{2} \log(|\Sigma|) \) into Eq. (S1–5), \( \Omega \) is canonical parameter using backward mapping. We get \( \Omega = (\mathbb{E}_\theta[X, XX^T] - \mathbb{E}_\theta[X] \mathbb{E}_\theta[X]^T)^{-1} = \Sigma^{-1} \), which can be inferred by the estimated covariance matrix.

S2 Appendix: Method and Optimization

**More about Proximal Optimization:** The proximal algorithm only needs to calculate the proximity operator of the parameters to be optimized. The proximity operator in proximal algorithms is defined as:

\[
\text{prox}_{\gamma f}(x) = \arg\min_{y} (f(y) + (\frac{1}{2\gamma} ||x - y||_2^2)). \tag{S2–1}
\]

The benefit of the proximal algorithm is that many proximity operators are entry-wise operators for the
targeted parameters. The parallel proximal (initially called proximity splitting) algorithm \[3\] belongs to the general family of distributed convex optimization that optimizes in such a way that each term (in this case, each proximity operator) can be handled by its own processing element, such as a thread or processor.

More about four proximity operators for CPU implementation of FASJEM-G: In the following, we denote \( x = \Omega_{\text{tot}}, a = \Sigma_{\text{tot}} \) and \( g \in \mathcal{G} \) to simply notations. Eq. (S:2–2) and Eq. (S:2–4) are entry-wise operators and Eq. (S:2–3) and Eq. (S:2–5) are group entry-wise. Group entry-wise means in calculation, the operator can compute each group of entries independently from other groups. Entry-wise means the calculation of each entry is only related to itself. The optimization process of Algorithm 1 iterating among four proximal operators is visualized by Figure S:1.

For \( f_1(\cdot) = \| \cdot \|_1 \).

\[
\text{prox}_{\gamma f_1}(x) = \text{prox}_{\gamma \| \cdot \|_1}(x) = \begin{cases} x^{(i)}_{j,k} - \gamma, & x^{(i)}_{j,k} > \gamma \\ 0, & \|x^{(i)}_{j,k}\| \leq \gamma \\ x^{(i)}_{j,k} + \gamma, & x^{(i)}_{j,k} < -\gamma \end{cases} \tag{S:2–2}
\]

Eq. (S:2–2) is the closed form solution of Eq. (S:2–1) when \( f = \| \cdot \|_1 \). Here \( j, k = 1, \ldots, p \) and \( i = 1, \ldots, K \). This is an entry-wise operator (i.e., the calculation of each entry is only related to itself).

Similarly, \( f_2(\cdot) = \| \cdot \|_2 \).

\[
\text{prox}_{\gamma f_2}(x_g) = \text{prox}_{\gamma \| \cdot \|_2}(x_g) = \begin{cases} x_g - \frac{\gamma}{\|x_g\|_2} \cdot x_g, & \|x_g\|_2 > \gamma \\ 0, & \|x_g\|_2 \leq \gamma \end{cases} \tag{S:2–3}
\]

Here \( g \in \mathcal{G} \). This is a group entry-wise operator (computing a group of entries is not related to other groups).

\( f_3(\cdot) \) and \( f_4(\cdot) \) include function forms of \( I_{f(\cdot) < D} \) and \( \text{prox}_{f(\cdot) < D} = \text{proj}_{f(\cdot) < D} \), where \( \text{proj}_C \) means the projection function to the convex set \( C \). We can obtain

\[
\text{prox}_{\gamma f_3}(x) = \text{proj}_{\| x - a \|_\infty \leq \lambda} \left( \begin{array}{c} x^{(i)}_{j,k} - a^{(i)}_{j,k} \leq \lambda \\ a^{(i)}_{j,k} + \lambda, x^{(i)}_{j,k} > a^{(i)}_{j,k} + \lambda \\ a^{(i)}_{j,k} - \lambda, x^{(i)}_{j,k} < a^{(i)}_{j,k} - \lambda \end{array} \right) \tag{S:2–4}
\]

where \( j, k = 1, \ldots, p \) and \( i = 1, \ldots, K \). This operator is entry-wise (i.e., only related to each entry of \( x \) and \( a \)).

\[
\text{prox}_{\gamma f_4}(x_g) = \text{proj}_{\| x - a \|_2^\gamma \leq \lambda} \left( \begin{array}{c} x_g - \frac{\gamma}{\|x_g - a_g\|_2} \cdot x_g, & \|x_g - a_g\|_2 \leq \lambda \\ \lambda \frac{x_g - a_g}{\|x_g - a_g\|_2^\gamma}, & \|x_g - a_g\|_2 > \lambda \end{array} \right) \tag{S:2–5}
\]

This operator is group entry-wise.

More about four proximity operators for GPU parallel implementation of FASJEM-G: The four proximity operators on GPU are summarized in Table 1. More details as followings:

For Eq. (S:2–2),

\[
\text{prox}_{\gamma f_1}(x) = \text{prox}_{\gamma \| \cdot \|_1}(x) = \max((x_{j,k}^{(i)} - \gamma), 0) + \min(0, (x_{j,k}^{(i)} + \gamma)) \tag{S:2–6}
\]

For Eq. (S:2–3),

\[
\text{prox}_{\gamma f_2}(x_g) = \text{prox}_{\gamma \| \cdot \|_2}(x_g) = x_g \max((1 - \frac{\gamma}{\|x_g\|_2}), 0) \tag{S:2–7}
\]

For Eq. (S:2–4),

\[
\text{prox}_{\gamma f_3}(x) = \text{proj}_{\| x - a \|_\infty \leq \lambda} = \min(\max(x_{j,k}^{(i)} - a_{j,k}^{(i)} - \lambda, \lambda) + a_{j,k}^{(i)}) \tag{S:2–8}
\]

For Eq. (S:2–5),

\[
\text{prox}_{\gamma f_4}(x_g) = \text{proj}_{\| x - a \|_2^\gamma \leq \lambda} = \max(\frac{\lambda}{\|x_g - a_g\|_2}, 1)(x_g - a_g) + a_g \tag{S:2–9}
\]

Here \( j, k = 1, \ldots, p \), \( i = 1, \ldots, K \) and \( g \in \mathcal{G} \).

More about Q-linearly Convergence of Optimization: The proposed optimization is a first-order method. Based on the recent study \[4\], the optimization sequence \( \{ \Omega^k \} \) for \( i = 1 \) to \( t \) iteration converges Q-linearly. Q-linearly means:

\[
\limsup_{k \to \infty} \frac{\|\Omega^{k+1} - \Omega^*\|}{\|\Omega^k - \Omega^*\|} \leq \rho \tag{S:2–10}
\]
Appendix: Related previous studies using elementary based estimators

Related previous studies based on elementary estimators are summarized in Table S:1.

Appendix: More about Experimental Setting and Baselines

Hyperparameter tuning: We have tried BIC method (used in [5]) for choosing the tuning parameter $\lambda_n$. As pointed out by ([6], [7] and [8]), the BIC or AIC method may not work well for the high-dimensional case. Therefore we have skipped adding the results from BIC or AIC.

In our experiments, we compare our model with the baselines by varying the same set of the tuning parameters.

Baseline: Recent literature [9] shows that the single sGGM has a close form solution through the EE estimator (i.e., no iteration). It is not fair to compare our estimator to such a closed-form sGGM estimator in terms of the speed or memory usage. Therefore we don’t include the single sGGM as a baseline.

Real World Experiments: We also tried FASJEM-I and JGL-groupinf on the three datasets. No matched interactions were found in one dataset. Therefore, we omit the results.

Appendix: More Experimental Results from Simulated Data

Figure S:3 represents a comparison between the single-task EE estimator for sGGM and GLasso estimator. We choose the $\Omega_i$ in the random graph model as the true graph. We obtain the two subfigures by varying $p$ in a set of {100, 200, 300, 400, 500}. The left subfigure is “AUC vs. $p$ (number of features)” while the right subfigure is “Time vs. $p$ (number of features)”. Figure S:3 shows that the elementary estimator has achieved similar performance of GLasso among different $p$ while the computation time of EE is much less than the GLasso.

Experiments on Real-world Datasets

We apply FASJEM-G and JGL-group on four different real-world datasets: (1) the breast/colon cancer data [10] (with 2 cell types and 104 samples, each having 22283 features); (2) Crohn’s disease data [11] (with 3 cell types, 127 samples and 22283 features), (3) the myeloma and bone lesions data set [12] (with 2 cell types, 173 samples and 12625 features) and (4) Encode project dataset [13] (with 3 cell types, 25185 samples and 27 features). For the first three datasets, we select its top 500 features based on the variance of the variables. After obtaining estimated dependency networks, we compare all methods using two major existing databases [14, 15] archiving known gene interactions. The number of known gene-gene interactions predicted by each method has been shown as bar graphs in Figure S:4. These graphs clearly show that FASJEM-G outperforms JGL-group on all three datasets and across all cell conditions within each of the three datasets. This leads us to believe that the proposed FASJEM-G is very promising for identifying variable interactions in a wider range of applications as well.
Table S:1: Two categories of relevant studies differ over learning based on “penalized log-likelihood” or learning based on “elementary estimator”

<table>
<thead>
<tr>
<th>Problems</th>
<th>Penalized Likelihood</th>
<th>Elementary estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>High dimensional linear regression</td>
<td>Lasso: argmin $\beta$ $</td>
<td>Y - \beta X</td>
</tr>
<tr>
<td>sparse Gaussian Graphical Model</td>
<td>gLasso: $\argmin_{\Omega \geq 0} -\logdet(\Omega) + &lt;\Omega, \Sigma&gt; + \lambda</td>
<td>\Omega</td>
</tr>
<tr>
<td>Multi-task sGGM</td>
<td>Different Choices for Penalty $R^*$ $\argmin_{\Omega \geq 0} \sum_i (-L(\Omega_{tot}) + \lambda_i \sum_i</td>
<td></td>
</tr>
</tbody>
</table>

Figure S:4: Compare predicted dependencies among genes or proteins using existing databases [14, 15] with known interactions (biologically validated) in human. The number of matches among predicted interactions and known interactions is shown as bar lines.

S:7 Appendix: More about the theoretical error bounds

Background–error bound for elementary estimator: For proving the error bounds, we first briefly review the error bound of a single-task EE-based model using the unified framework [2]. The single task-EE follows the general formulation:

$$\argmin_{\theta} R(\theta)$$
subject to $R^*(\hat{\theta}_n - \theta) \leq \lambda_n$ (S:7–1)

where $R(\cdot)$ is the $\ell_1$ regularization function and $\hat{\theta}_n$ is the backforward mapping for $\theta$.

Following the unified framework [2], we first decompose the parameter space into a subspace pair $(\mathcal{M}, \mathcal{M}^\perp)$, where $\mathcal{M}$ is the closure of $\mathcal{M}$. Here $\mathcal{M}$ is the model subspace that typically has a much lower dimension than the original high-dimensional space. $\mathcal{M}^\perp$ is the perturbation subspace of parameters. For further proofs, we assume the regularization function in Eq. (S:7–1) is decomposable w.r.t the subspace pair $(\mathcal{M}, \mathcal{M}^\perp)$.

(C1) $R(u + v) = R(u) + R(v)$, $\forall u \in \mathcal{M}, \forall v \in \mathcal{M}^\perp$.

[2] shows that most regularization norms are decomposable corresponding to a certain subspace pair.

Definition S:7.1. A term subspace compatibility constant is defined as $\Psi(\mathcal{M}, |\cdot|) := \sup_{\mu \in \mathcal{M}\setminus\{0\}} \frac{R(u)}{|\mu|}$, which captures the relative value between the error norm $|\cdot|$ and the regularization function $R(\cdot)$.

For simplicity, we assume there exists a true parameter $\theta^*$ which has the exact structure w.r.t a certain subspace pair. That is:

(C2) $\exists$ a subspace pair $(\mathcal{M}, \mathcal{M}^\perp)$ such that the true parameter satisfies proj$_{\mathcal{M}^\perp}(\theta^*) = 0$

Then we have the following theorem.

Theorem S:7.2. Suppose the regularization function in Eq. (S:7–1) satisfies condition (C1), the true parameter of Eq. (S:7–1) satisfies condition (C2), and $\lambda_n$ satisfies that $\lambda_n \geq R^*(\hat{\theta} - \theta^*)$. Then, the optimal solution $\hat{\theta}$ of Eq. (S:7–1) satisfies:

\begin{align*}
R^*(\hat{\theta} - \theta^*) & \leq 2\lambda_n \quad (S:7–2) \\
||\hat{\theta} - \theta^*||_2 & \leq 4\lambda_n \Psi(\mathcal{M}) \quad (S:7–3) \\
R(\hat{\theta} - \theta^*) & \leq 8\lambda_n \Psi(\mathcal{M})^2 \quad (S:7–4)
\end{align*}

S:8 Proof

Proof of Theorem (S:7.2)

Proof. Let $\Delta := \hat{\theta} - \theta^*$ be the error vector that we are interested in.

\begin{align*}
R^*(\hat{\theta} - \theta^*) & = R^*(\hat{\theta} - \hat{\theta}_n + \hat{\theta}_n - \theta^*) \\
& \leq R^*(\hat{\theta}_n - \hat{\theta}) + R^*(\hat{\theta}_n - \theta^*) \leq 2\lambda_n \quad (S:8–1)
\end{align*}
where \( \Psi(\bar{\mathcal{M}}) = 0 \), and the decomposability of \( \mathcal{R} \) with respect to \((\mathcal{M}, \mathcal{M}^\perp)\)

\[
\mathcal{R}(\theta^*) = |\mathcal{R}(\theta^*) + \mathcal{R}[\Pi_{\mathcal{M}^\perp}(\Delta)] - \mathcal{R}[\Pi_{\mathcal{M}^\perp}(\Delta)]|
\]

\[
\leq |\mathcal{R}(\theta^* + \Pi_{\mathcal{M}^\perp}(\Delta) + \Pi_{\mathcal{M}^\perp}(\Delta)) + \mathcal{R}[\Pi_{\mathcal{M}^\perp}(\Delta)] - \mathcal{R}[\Pi_{\mathcal{M}^\perp}(\Delta)]|.
\]

\[
= |\mathcal{R}(\theta^* + \Delta) + \mathcal{R}[\Pi_{\mathcal{M}^\perp}(\Delta)] - \mathcal{R}[\Pi_{\mathcal{M}^\perp}(\Delta)]|.
\]

(8.8-2)

Here, the inequality holds by the triangle inequality of norm. Since Eq. (8.7-1) minimizes \( \mathcal{R}(\hat{\theta}) \), we have \( \mathcal{R}(\theta^* + \Delta) = \mathcal{R}(\hat{\theta}) \leq \mathcal{R}(\theta^*) \). Combining this inequality with Eq. (8.8-2), we have:

\[
\mathcal{R}[\Pi_{\mathcal{M}^\perp}(\Delta)] \leq \mathcal{R}[\Pi_{\mathcal{M}^\perp}(\Delta)].
\]

(8.8-3)

Moreover, by Hölder’s inequality and the decomposability of \( \mathcal{R}(\cdot) \), we have:

\[
||\Delta||_2^2 = |\Delta, \Delta| \leq \mathcal{R}^*(\Delta) \mathcal{R}(\Delta) \leq 2\lambda_n \mathcal{R}(\Delta)
\]

\[
= 2\lambda_n [\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\Delta))] + \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\Delta))] \leq 4\lambda_n \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\Delta))
\]

\[
\leq 4\lambda_n \Psi(\bar{\mathcal{M}})||\Pi_{\mathcal{M}^\perp}(\Delta)||_2
\]

(8.8-4)

where \( \Psi(\bar{\mathcal{M}}) \) is a simple notation for \( \Psi(\bar{\mathcal{M}}, ||\cdot||_2) \).

Since the projection operator is defined in terms of \( ||\cdot||_2 \) norm, it is non-expansive: \( ||\Pi_{\mathcal{M}^\perp}(\Delta)||_2 \leq ||\Delta||_2 \). Therefore, by Eq. (8.8-4), we have:

\[
||\Pi_{\mathcal{M}^\perp}(\Delta)||_2 \leq 4\lambda_n \Psi(\bar{\mathcal{M}}).
\]

(8.8-5)

and plugging it back to Eq. (8.8-4) yields the error bound Eq. (8.8-3).

Finally, Eq. (8.7-4) is straightforward from Eq. (8.8-3) and Eq. (8.8-5).

\[
\mathcal{R}(\Delta) \leq 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\Delta))
\]

\[
\leq 2\Psi(\bar{\mathcal{M}})||\Pi_{\mathcal{M}^\perp}(\Delta)||_2 \leq 8\lambda_n \Psi(\bar{\mathcal{M}})^2.
\]

(8.8-6)

\[\square\]

**Proof of Theorem (5.3)**

*Proof.* In this proof, we consider the matrix parameter such as the covariance. \( I = \{1, 2\} \) in the following contents. Basically, the Frobenius norm can be simply replaced by \( \ell_2 \) norm for the vector parameters. Let \( \Delta_i := \widehat{\theta}_i - \theta^*_i \), and \( \Delta = \theta - \theta^* = \Sigma_{i \in I} \Delta_i \). The error bound Eq. (5.3) can be easily shown from the assumption in the statement with the constraint of Eq. (5.2).

For every \( i \in I \),

\[
\mathcal{R}_i(\Delta_i) = \mathcal{R}_i^*(\hat{\theta} - \theta^*) = \mathcal{R}_i^*(\hat{\theta} - \theta_n + \theta_n - \theta^*)
\]

\[
\leq \mathcal{R}_i^*(\hat{\theta}_n - \hat{\theta}_n) + \mathcal{R}_i^*(\hat{\theta}_n - \theta^*) \leq 2\lambda_i.
\]

(8.8-7)

By the similar reasoning as in Eq. (8.8-2) with the fact that \( \Pi_{\mathcal{M}^\perp_i}(\theta^*_i) = 0 \) in \( \text{C3} \), and the decomposability of \( \mathcal{R}_i \) with respect to \((\mathcal{M}_i, \mathcal{M}_i^\perp)\), we have:

\[
\mathcal{R}_i(\theta^*_i) \leq \mathcal{R}_i(\theta^*_i + \Delta_i) + \mathcal{R}_i(\Pi_{\mathcal{M}^\perp_i}(\Delta_i))
\]

\[
- \mathcal{R}_i(\Pi_{\mathcal{M}^\perp_i}(\Delta_i)),
\]

(8.8-8)

Since \( \{\hat{\theta}_i\}_i \) minimizes the objective function of Eq. (5.2),

\[
\sum_{i \in I} \lambda_i \mathcal{R}_i(\hat{\theta}_i) \leq \sum_{i \in I} \lambda_i \{\mathcal{R}_i(\theta^*_i + \Delta_i)
\]

\[
- \mathcal{R}_i(\Pi_{\mathcal{M}^\perp_i}(\Delta_i))\},
\]

(8.8-9)

Which implies

\[
\sum_{i \in I} \lambda_i \mathcal{R}_i(\Pi_{\mathcal{M}^\perp_i}(\Delta_i)) \leq \sum_{i \in I} \lambda_i \mathcal{R}_i(\Pi_{\mathcal{M}^\perp_i}(\Delta_i))
\]

(8.8-10)

Now, for each structure \( i \in I \), we have an application for Hölder’s inequality: \( ||\Delta_i||_F \leq \mathcal{R}_i^*(\Delta_i)/\mathcal{R}_i(\Delta_i) \leq 2\lambda_i/\mathcal{R}_i(\Delta_i) \) where the notation \( \langle A, B \rangle \) denotes the trace inner product, \( \text{trace}(A^T B) = M_{ij} \Sigma_j A_{ij} B_{ij} \), and we use the pre-computed bound in Eq. (8.8-7). Then, the Frobenius error \( ||\Delta||_F \) can be upper-bounded as follows:

\[
||\Delta||_F^2 = \sum_{i \in I} \langle \Delta_i, \Delta_i \rangle \leq \sum_{i \in I} ||\Delta_i||^2
\]

\[
\leq 2 \sum_{i \in I} \lambda_i \mathcal{R}_i(\Delta_i) \leq 2 \sum_{i \in I} \{\lambda_i \mathcal{R}_i(\Pi_{\mathcal{M}^\perp_i}(\Delta_i)) + \lambda_i \mathcal{R}_i(\Pi_{\mathcal{M}^\perp_i}(\Delta_i))\}
\]

\[
\leq 4 \sum_{i \in I} \lambda_i \Psi(\bar{\mathcal{M}})||\Pi_{\mathcal{M}^\perp_i}(\Delta_i)||_F
\]

(8.8-11)
where $\Psi(\mathcal{M}_i)$ denotes the compatibility constant of space $\mathcal{M}_i$ with respect to the Frobenius norm: $\Psi(\mathcal{M}_i, \| \cdot \|_F)$.

Here, we define a key notation in the error bound:

$$\Phi := \max_{i \in I} \lambda_i \Psi(\mathcal{M}_i).$$  \hfill (S:8–12)

Armed with this notation, Eq. (S:8–11) can be written

$$\|\Delta\|_F^2 \leq 4\Phi \sum_{i \in I} \|\Pi_{\mathcal{M}_i}(\Delta_i)\|_F^2$$ \hfill (S:8–13)

At this point, we directly appeal to the result in Proposition 2 of [16] with a small modification:

**Proposition 4.** Suppose that the structural incoherence condition (C4) as well as the condition (C3) hold. Then, we have

$$2 \left| \sum_{i < j} \langle \Delta_i, \Delta_j \rangle \right| \leq \frac{1}{2} \sum_{i \in I} \|\Delta_i\|_F^2.$$ \hfill (S:8–14)

By this proposition, we have

$$\sum_{i \in I} \|\Delta_i\|_F^2 \leq \|\Delta\|_F^2 + 2 \left| \sum_{i < j} \langle \Delta_i, \Delta_j \rangle \right|$$

$$\leq \|\Delta\|_F^2 + \frac{1}{2} \sum_{i \in I} \|\Delta_i\|_F^2,$$ \hfill (S:8–15)

which implies $\sum_{i \in I} \|\Delta_i\|_F^2 \leq 2\|\Delta\|_F^2$.

Moreover, since the projection operator is defined in terms of the Frobenius norm, it is non-expansive for all $i : \|\Pi_{\mathcal{M}_i}(\Delta_i)\|_F \leq \|\Delta_i\|_F$. Hence, we finally obtain:

$$(\sum_{i \in I} \|\Pi_{\mathcal{M}_i}(\Delta_i)\|_F^2)^2 \leq (\sum_{i \in I} \|\Delta_i\|_F^2)^2$$

$$\leq |I| \sum_{i \in I} \|\Delta_i\|_F^2 \leq 8|I|\Phi \sum_{i \in I} \|\Pi_{\mathcal{M}_i}(\Delta_i)\|_F$$ \hfill (S:8–16)

and therefore,

$$\sum_{i \in I} \|\Pi_{\mathcal{M}_i}(\Delta_i)\|_F \leq 8|I|\Phi$$ \hfill (S:8–17)

The Frobenius norm error bound Eq. (S:8–18) can be derived by plugging Eq. (S:8–17) back into Eq. (S:8–18):}

$$\|\Delta\|_F^2 \leq 32|I|\Phi^2.$$ \hfill (S:8–18)

Therefore, we have

$$\|\Delta\|_F \leq 8\Phi$$ \hfill (S:8–19)

Which is exactly Eq. (5.5)

The proof of the final error bound Eq. (5.4) is straightforward from Eq. (S:8–10) and Eq. (S:8–17) as follows: for each fixed $i \in I$,

$$R_i(\Delta_i)$$

$$\leq \frac{1}{\lambda_i} \{ \lambda_i R_i[\Pi_{\mathcal{M}_i}(\Delta_i)] + \lambda_i R_i[\Pi_{\mathcal{M}_i}^T(\Delta_i)] \}$$

$$\leq \frac{1}{\lambda_i} \{ \lambda_i R_i[\Pi_{\mathcal{M}_i}(\Delta_i)] + \sum_{j \in I} \lambda_j R_j[\Pi_{\mathcal{M}_j}(\Delta_j)] \}$$

$$\leq \frac{2}{\lambda_i} \sum_{j \in I} \lambda_j R_j[\Pi_{\mathcal{M}_j}(\Delta_j)]$$

$$\leq \frac{2}{\lambda_i} \sum_{j \in I} \lambda_j \Psi(\mathcal{M}_j) \|\Pi_{\mathcal{M}_j}(\Delta_j)\|_F$$

$$\leq \frac{2\Phi}{\lambda_i} \sum_{j \in I} \|\Pi_{\mathcal{M}_j}(\Delta_j)\|_F \leq \frac{16|I|\Phi^2}{\lambda_i} = \frac{32\Phi^2}{\lambda_i}$$ \hfill (S:8–20)

which completes the proof. \hfill \Box

**Proof of Theorem (5.4)**

**Proof.** Since $\lambda_n > \lambda'_n$ and $\sqrt{s} > \sqrt{s_n}$, We have that $\lambda_n \sqrt{s} > \lambda'_n \sqrt{s_n}$.

By Theorem (5.3),

$$||\Omega_{tot} - \Omega_{tot}'\|_F \leq 8 \max(\lambda_n \sqrt{s}, \lambda'_n \sqrt{s_n}) \leq 8 \sqrt{s} \lambda_n.$$ \hfill \Box

**S.8.1 Useful lemma(s)**

**Lemma S:8.1.** (Theorem 1 of [17]). Let $\delta$ be $\max_{ij} \| \frac{X^T X}{n} \|_{ij} - \Sigma_{ij}$. Suppose that $\nu > 2\delta$. Then, under the conditions (C-Sparse$\Sigma$), and as $p_n(\cdot)$ is a soft-thresholding function, we can deterministically guarantee that the spectral norm of error is bounded as follows:

$$\|T_v(\widehat{\Sigma}) - \Sigma\|_\infty \leq 5\nu^{1-\frac{1}{2}}c_0(p) + 3\nu^{1-\frac{1}{2}}c_0(p)\delta$$ \hfill (S:8–21)

**Lemma S:8.2.** (Lemma 1 of [18]). Let $\mathcal{A}$ be the event that

$$\left\| \frac{X^T X}{n} - \Sigma \right\|_\infty \leq 8(\max_{i} \Sigma_{ii}) \sqrt{\frac{10\rho \log p'}{n}}$$ \hfill (S:8–22)

where $p' := \max n, p$ and $\tau$ is any constant greater than 2. Suppose that the design matrix $X$ is i.i.d. sampled
from $\Sigma$-Gaussian ensemble with $n \geq 40\max_i \Sigma_{ii}$. Then, the probability of event $A$ occurring is at least $1 - 4/p^{\tau-2}$.

**Proof of Corollary (5.5)**

Proof. In the following proof, we re-define the following two notations: $\Sigma_{\text{tot}} := \begin{pmatrix} \Sigma^{(1)} & 0 & \cdots & 0 \\ 0 & \Sigma^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma^{(K)} \end{pmatrix}$ and $\Omega_{\text{tot}} := \begin{pmatrix} \Omega^{(1)} & 0 & \cdots & 0 \\ 0 & \Omega^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Omega^{(K)} \end{pmatrix}$

The condition (C-Sparse$\Sigma$) and condition (C-MinInf$\Sigma$) also hold for $\Omega_{\text{tot}}$ and $\Sigma_{\text{tot}}$. In order to utilize Theorem (5.4) for this specific case, we only need to show that $\left|\|T_{\nu}(\widehat{\Sigma}_{\text{tot}})\|^{-1}\|_{\infty} \leq \lambda_n$ for the setting of $\lambda_n$ in the statement:

\[
\left|\|T_{\nu}(\widehat{\Sigma}_{\text{tot}})\|^{-1}\|_{\infty} = \left|\|T_{\nu}(\widehat{\Sigma}_{\text{tot}})\|^{-1}(T_{\nu}(\widehat{\Sigma}_{\text{tot}})\Omega_{\text{tot}}^* - I)\|_{\infty}
\right.
\]
\[
\leq \left|\|T_{\nu}(\widehat{\Sigma}_{\text{tot}})w\|\|_{\infty} \|T_{\nu}(\widehat{\Sigma}_{\text{tot}})\Omega_{\text{tot}}^* - I\|_{\infty}
\right.
\]
\[
= \left|\|T_{\nu}(\widehat{\Sigma}_{\text{tot}})\|^{-1}\|_{\infty} \|\Omega_{\text{tot}}^*(T_{\nu}(\widehat{\Sigma}_{\text{tot}}) - \Sigma_{\text{tot}})\|_{\infty}
\right.
\]
\[
\leq \left|\|T_{\nu}(\widehat{\Sigma}_{\text{tot}})\|^{-1}\|_{\infty} \|\Omega_{\text{tot}}\|_{\infty} \|T_{\nu}(\widehat{\Sigma}_{\text{tot}}) - \Sigma_{\text{tot}}\|_{\infty}
\right.
\]

(S8–23)

We first compute the upper bound of $\left|\|T_{\nu}(\widehat{\Sigma}_{\text{tot}})\|^{-1}\|_{\infty}$. By the selection $\nu$ in the statement, Lemma (S8.1) and Lemma (S8.2) hold with probability at least $1 - 4/p^{\tau-2}$. Armed with Eq. (S8–21), we use the triangle inequality of norm and the condition (C-Sparse$\Sigma$) for any $w$,

\[
\|T_{\nu}(\widehat{\Sigma}_{\text{tot}})w\|_{\infty} = \|T_{\nu}(\widehat{\Sigma}_{\text{tot}})w - \Sigma w + \Sigma w\|_{\infty}
\]
\[
\geq \|\Sigma w\|_{\infty} - \|T_{\nu}(\widehat{\Sigma}_{\text{tot}} - \Sigma)w\|_{\infty}
\]
\[
\geq \kappa_2\|w\|_{\infty} - \|T_{\nu}(\widehat{\Sigma}_{\text{tot}} - \Sigma)w\|_{\infty}
\]
\[
\geq (\kappa_2 - \|T_{\nu}(\widehat{\Sigma}_{\text{tot}} - \Sigma)w\|_{\infty})\|w\|_{\infty}
\]

(S8–24)

Where the second inequality uses the condition (C-Sparse$\Sigma$). Now, by Lemma (S8.1) with the selection of $\nu$, we have

\[
\left|\|T_{\nu}(\widehat{\Sigma}_{\text{tot}})\|^{-1}\|_{\infty} \leq c_1(\frac{\log p^{'}}{n_{\text{tot}}})^{(1-q)/2}c_0(p)
\]

(S8–25)

where $c_1$ is a constant related only on $\tau$ and $\max_i \Sigma_{ii}$. Specifically, it is defined as $6.5(16\max_i \Sigma_{ii})\sqrt{10}/(\tau r^2)$. Hence, as long as $n_{\text{tot}} > (\frac{2\log(p^{'})}{\kappa_2}) \log p$ as stated, so that $\left|\|T_{\nu}(\widehat{\Sigma}_{\text{tot}}) - \Sigma\|\|_{\infty} \leq \frac{\kappa_2}{2}$, we can conclude that $\|T_{\nu}(\widehat{\Sigma}_{\text{tot}})w\|_{\infty} \geq \frac{\kappa_2}{2}\|w\|_{\infty}$, which implies $\left|\|T_{\nu}(\widehat{\Sigma}_{\text{tot}})\|^{-1}\|_{\infty} \leq \frac{2}{\kappa_2}$.

The remaining term in Eq. (S8–23) is $\|T_{\nu}(\widehat{\Sigma}_{\text{tot}}) - \Sigma_{\text{tot}}\|_{\infty}$: $\|T_{\nu}(\widehat{\Sigma}_{\text{tot}}) - \Sigma_{\text{tot}}\|_{\infty} \leq \|T_{\nu}(\widehat{\Sigma}_{\text{tot}}) - \Sigma_{\text{tot}}\|_{\infty} + \|\Sigma_{\text{tot}} - \Sigma_{\text{tot}}\|_{\infty}$. By construction of $T_{\nu}(\cdot)$ in (C-Thresh) and by Lemma (S8.2), we can confirm that $\|T_{\nu}(\widehat{\Sigma}_{\text{tot}}) - \Sigma_{\text{tot}}\|_{\infty}$ as well as $\|\Sigma_{\text{tot}} - \Sigma_{\text{tot}}\|_{\infty}$ can be upper-bounded by $\nu$.

By combining all together, we can confirm that the selection of $\lambda_n$ satisfies the requirement of Theorem (5.4), which completes the proof.

**References**


