Appendix: A Fast and Scalable Joint Estimator for Learning Multiple Related Sparse Gaussian Graphical Models

Beilun Wang
University of Virginia

Ji Gao
University of Virginia

Yanjun Qi
University of Virginia

S:1 Appendix: Backward mapping for M-Estimator

The graphical model MLE can be expressed as a backward mapping[1] 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:

$$\mathbb{P}(X;\theta) = h(X)\exp\{<\theta, \phi(\theta)> - A(\theta)\} \quad \text{(S:1–1)}$$

Where $\theta \in \Theta \subset \mathbb{R}^d$ is the canonical parameter to be estimated and Θ denotes the parameter space, $\phi(X)$ denotes the sufficient statistics with a feature mapping function $\phi : \mathbb{R}^p \to \mathbb{R}^d$, 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:

$$\mathcal{M} = \{\nu | \exists p \text{ is a distribution s.t. } \mathbb{E}_p[\phi(X)] = \nu\} \quad \text{(S:1–2)}$$

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

$$\mathcal{A} : \text{H} \to \mathcal{M} \quad \text{(S:1–3)}$$

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:

$$\mathcal{A}^* : \mathcal{M}^0 \to \text{H} \quad \text{(S:1–4)}$$

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

$$\mathcal{A}^* : \nu(\theta) \to \theta = \nabla A^*(\nu(\theta)). \quad \text{(S:1–5)}$$

Where $A^*(\nu(\theta)) = \sup_{\theta \in \text{H}} <\theta, \nu(\theta)> - A(\theta)$.

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

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 θ contains entry Σ^{-1}, when estimating sGGM, we need to use the backward mapping:

For the case of Gaussian distribution,

$$\begin{aligned} \theta &= (\Sigma^{-1}\mu, -\frac{1}{2}\Sigma^{-1}) = \mathcal{A}^*(\nu) = \nabla A^*(\nu) \\ &= ((\mathbb{E}_\theta[XX^T] - \mathbb{E}_\theta[X]\mathbb{E}_\theta[X]^T)^{-1}\mathbb{E}_\theta[X], \\ &\quad -\frac{1}{2}(\mathbb{E}_\theta[XX^T] - \mathbb{E}_\theta[X]\mathbb{E}_\theta[X]^T)^{-1}). \end{aligned} \quad \text{(S:1–6)}$$

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

S:2 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:

$$\text{prox}_{\gamma f}(x) = \underset{y}{\text{argmin}}(f(y) + (\frac{1}{2\gamma}||x-y||_2^2)). \quad \text{(S:2–1)}$$

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.

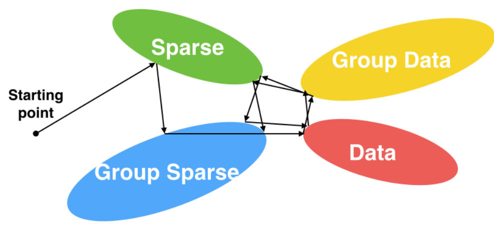

Figure S:1: A simple figure to show how our optimization method works. Our optimization approach is a method with linear convergence rate in finding the optimal point. It considers four properties : (1) information from the raw data; (2) information from the group data; (3)sparsity property; (4) group sparsity property.

More about four proximity operators for CPU implementation of FASJEM-G: In the following, we denote $x = \Omega_{tot}$, $a = \Sigma_{tot}$ 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, $f_2(\cdot) = ||\cdot||_{\mathcal{G},2}$

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

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 $\mathcal{I}_{f(\cdot)<D}$ and $\text{prox}_{\mathcal{I}_{\{f(\cdot)<D\}}} = \text{proj}_{\{f(\cdot)<D\}}$, where proj_C means the projection function to the convex set C. We can obtain

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

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).

$$\text{prox}_{\gamma f_4}(x_g) = \text{proj}_{||x-a||^*_{\mathcal{G},2} \leq \lambda} = \begin{cases} x_g, \ ||x_g - a_g||_2 \leq \lambda \\ \lambda \frac{x_g - a_g}{||x_g - a_g||_2} + a_g, \ ||x_g - a_g||_2 > \lambda \end{cases} \tag{S:2–5}$$

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 following:

For Eq. (S:2–2),

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

For Eq. (S:2–3)

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

For Eq. (S:2–4)

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

For Eq. (S:2–5)

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

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 $\{\Omega^i\}$(for $i = 1$ to t iteration) converges Q-linearly. Q-linearly means:

$$\limsup_{k \to \infty} \frac{||\Omega^{k+1} - \Omega^*||}{||\Omega^k - \Omega^*||} \leq \rho \tag{S:2–10}$$

S:3 Appendix: Related previous studies using elementary based estimators

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

S:4 Appendix: More about Experimental Setting and Baselines

Hyperparameter tuning: We have tried BIC method (used in [5]) for choosing the tuning parameter λ_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.

S:5 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.

S:6 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.

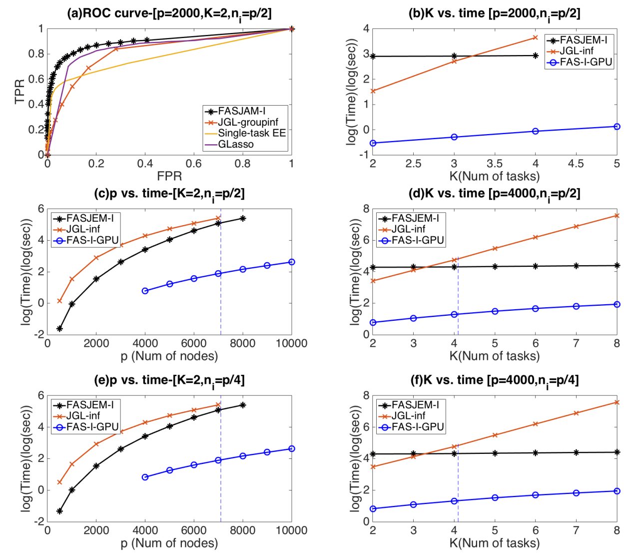

Figure S:2: Comparison between FASJEM-I and JGL-groupinf using accuracy, speed and memory capacity. (a) FPR-TPR curves of two methods on the simulated dataset using Random Graph Model when $p = 2000$ and $K = 2$. (c) and (e) Time versus p(the number of variables) curves from FASJEM-G, JGL-group and FASJEM-I's GPU implementation. (c) uses $n_i = p/2$ and (e)$n_i = p/4$ (b), (d) and (f) include the time versus K(the number of tasks) curves for two methods plus FASJEM-I-GPU. (b) uses $p = 2000$ and $n_i = p/2$, (d) uses $p = 4000$ and $n_i = p/2$ and (f) uses $p = 4000$ and $n_i = p/4$.

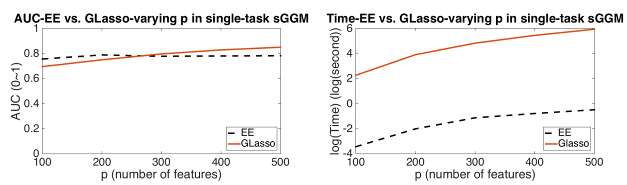

Figure S:3: Comparison between elementary estimator for sGGM and GLasso for single-task sGGM. The left figure is the curve of AUC number by varying p. The number of sample $n = p/2$. The right figure is the curve of computation time by varying p. Other settings are the same as the left one. Clearly, elementary estimator has the similar accuracy performance as GLasso but is much faster and scalable than it.

Table S:1: Two categories of relevant studies differ over learning based on “penalized log-likelihood" or learning based on“elementary estimator"

Problems	Penalized Likelihood	Elementary estimator
High dimensional linear regression	Lasso: $\underset{\beta}{\text{argmin}}\, \lvert Y-\beta X\rvert_F+\lambda\lvert\beta\rvert_1$	$\underset{\beta}{\text{argmin}}\, \lvert\beta\rvert_1$ subject to : $\lvert\beta-(X^TX+\epsilon I)^{-1}X^Ty\rvert_\infty \leq \lambda_n$
sparse Gaussian Graphical Model	gLasso: $\underset{\Omega\geq 0}{\text{argmin}} -logdet(\Omega)+<\Omega,\Sigma> +\lambda\lvert\Omega\rvert_1$	$\underset{\Omega\geq 0}{\text{argmin}}\, \lvert\Omega\rvert_1$ subject to: $\lvert\Omega-[T_v(\Sigma)]^{-1}\rvert_\infty \leq \lambda_n$
Multi-task sGGM	Different Choices for Penalty $\mathcal{R}'$ $\underset{\Omega>0}{\text{argmin}} \sum_i(-L(\Omega_{tot})+\lambda_1\sum_i \lvert\lvert\Omega^{(i)}\rvert\rvert_1 + \lambda_2\mathcal{R}'(\Omega_{tot})$	**Our method: FASJEM**

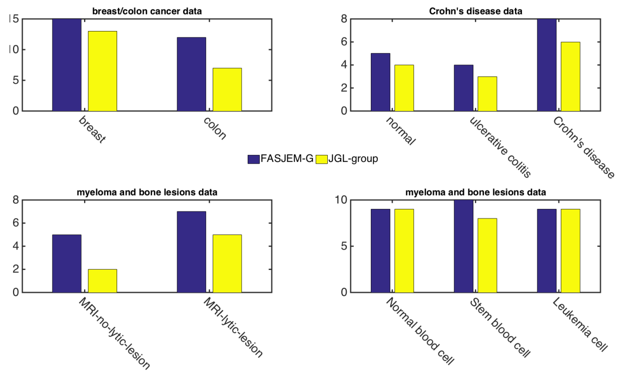

Figure S:4: Compare predicted dependencies among genes or proteins using existing databases [14, 15] with known interactions (biologically validated) in human. The number of matches among predicted interactions and known interactions is shown as bar lines.

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:

$$\begin{aligned}&\underset{\theta}{\text{argmin}}\mathcal{R}(\theta)\\ &\text{subject to:}\mathcal{R}^*(\widehat{\theta}_n-\theta)\leq\lambda_n\end{aligned} \tag{S:7–1}$$

where $\mathcal{R}(\cdot)$ is the ℓ_1 regularization function and $\widehat{\theta}_n$ is the backforward mapping for θ.

Following the unified framework [2], we first decompose the parameter space into a subspace pair$(\mathcal{M},\bar{\mathcal{M}}^\perp)$, where $\bar{\mathcal{M}}$ 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. $\bar{\mathcal{M}}^\perp$ 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 $(\mathcal{M},\bar{\mathcal{M}}^\perp)$.

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

[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* $\Psi(\mathcal{M},|\cdot|):=\sup_{u\in\mathcal{M}\setminus\{0\}}\frac{\mathcal{R}(u)}{|u|}$ *which captures the relative value between the error norm* $|\cdot|$ *and the regularization function* $\mathcal{R}(\cdot)$.

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

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

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* λ_n *satisfies that* $\lambda_n\geq\mathcal{R}^*(\widehat{\theta}-\theta^*)$*. Then, the optimal solution* $\widehat{\theta}$ *of Eq. (S:7–1) satisfies:*

$$\mathcal{R}^*(\widehat{\theta}-\theta^*)\leq 2\lambda_n \tag{S:7–2}$$

$$||\widehat{\theta}-\theta^*||_2\leq 4\lambda_n\Psi(\bar{\mathcal{M}}) \tag{S:7–3}$$

$$\mathcal{R}(\widehat{\theta}-\theta^*)\leq 8\lambda_n\Psi(\bar{\mathcal{M}})^2 \tag{S:7–4}$$

S:8 Proof

Proof of Theorem (S:7.2)

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

$$\begin{aligned}&\mathcal{R}^*(\widehat{\theta}-\theta^*)=\mathcal{R}^*(\widehat{\theta}-\widehat{\theta}_n+\widehat{\theta}_n-\theta^*)\\ &\leq\mathcal{R}^*(\widehat{\theta}_n-\widehat{\theta})+\mathcal{R}^*(\widehat{\theta}_n-\theta^*)\leq 2\lambda_n\end{aligned} \tag{S:8–1}$$

By the fact that $\theta^*_{\mathcal{M}^\perp} = 0$, and the decomposability of $\mathcal{R}$ with respect to $(\mathcal{M}, \bar{\mathcal{M}}^\perp)$

$$\begin{aligned}
&\mathcal{R}(\theta^*) \\
&= \mathcal{R}(\theta^*) + \mathcal{R}[\Pi_{\bar{\mathcal{M}}^\perp}(\Delta)] - \mathcal{R}[\Pi_{\bar{\mathcal{M}}^\perp}(\Delta)] \\
&= \mathcal{R}[\theta^* + \Pi_{\bar{\mathcal{M}}^\perp}(\Delta)] - \mathcal{R}[\Pi_{\bar{\mathcal{M}}^\perp}(\Delta)] \\
&\le \mathcal{R}[\theta^* + \Pi_{\bar{\mathcal{M}}^\perp}(\Delta) + \Pi_{\bar{\mathcal{M}}}(\Delta)] + \mathcal{R}[\Pi_{\bar{\mathcal{M}}}(\Delta)] \\
&- \mathcal{R}[\Pi_{\bar{\mathcal{M}}^\perp}(\Delta)] \\
&= \mathcal{R}[\theta^* + \Delta] + \mathcal{R}[\Pi_{\bar{\mathcal{M}}}(\Delta)] - \mathcal{R}[\Pi_{\bar{\mathcal{M}}^\perp}(\Delta)]
\end{aligned} \tag{S:8–2}$$

Here, the inequality holds by the triangle inequality of norm. Since Eq. (S:7–1) minimizes $\mathcal{R}(\widehat{\theta})$, we have $\mathcal{R}(\theta^* + \Delta) = \mathcal{R}(\widehat{\theta}) \le \mathcal{R}(\theta^*)$. Combining this inequality with Eq. (S:8–2), we have:

$$\mathcal{R}[\Pi_{\bar{\mathcal{M}}^\perp}(\Delta)] \le \mathcal{R}[\Pi_{\bar{\mathcal{M}}}(\Delta)] \tag{S:8–3}$$

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

$$\begin{aligned}
&||\Delta||_2^2 = \langle \Delta, \Delta \rangle \le \mathcal{R}^*(\Delta)\mathcal{R}(\Delta) \le 2\lambda_n \mathcal{R}(\Delta) \\
&= 2\lambda_n[\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta)) + \mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta))] \le 4\lambda_n \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta)) \\
&\le 4\lambda_n \Psi(\bar{\mathcal{M}}) ||\Pi_{\bar{\mathcal{M}}}(\Delta)||_2
\end{aligned} \tag{S:8–4}$$

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: $||\Pi_{\bar{\mathcal{M}}}(\Delta)||_2 \le ||\Delta||_2$. Therefore, by Eq. (S:8–4), we have:

$$||\Pi_{\bar{\mathcal{M}}}(\Delta)||_2 \le 4\lambda_n \Psi(\bar{\mathcal{M}}), \tag{S:8–5}$$

and plugging it back to Eq. (S:8–4) yields the error bound Eq. (S:7–3).

Finally, Eq. (S:7–4) is straightforward from Eq. (S:8–3) and Eq. (S:8–5).

$$\begin{aligned}
&\mathcal{R}(\Delta) \le 2\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta)) \\
&\le 2\Psi(\bar{\mathcal{M}}) ||\Pi_{\bar{\mathcal{M}}}(\Delta)||_2 \le 8\lambda_n \Psi(\bar{\mathcal{M}})^2.
\end{aligned} \tag{S:8–6}$$

□

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 ℓ_2 norm for the vector parameters. Let $\Delta_i := \widehat{\theta}_i - \theta^*_i$, and $\Delta = \widehat{\theta} - \theta^* = \Sigma_{i \in I} \Delta_i$. 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 $i \in I$,

$$\begin{aligned}
&\mathcal{R}^*_i(\Delta) = \mathcal{R}^*_i(\widehat{\theta} - \theta^*) = \mathcal{R}^*_i(\widehat{\theta} - \widehat{\theta}_n + \widehat{\theta}_n - \theta^*) \\
&\le \mathcal{R}^*_i(\widehat{\theta}_n - \widehat{\theta}) + \mathcal{R}^*_i(\widehat{\theta}_n - \theta^*) \le 2\lambda_i.
\end{aligned} \tag{S:8–7}$$

By the similar reasoning as in Eq. (S:8–2) with the fact that $\Pi_{\mathcal{M}_i^\perp}(\theta^*_i) = 0$ in **C3**, and the decomposability of $\mathcal{R}_i$ with respect to $(\mathcal{M}_i, \widehat{\mathcal{M}}_i^\perp)$, we have:

$$\begin{aligned}
\mathcal{R}_i(\theta^*_i) \le &\mathcal{R}_i[\theta^*_i + \Delta_i] + \mathcal{R}_i[\Pi_{\bar{\mathcal{M}}_i}(\Delta_i)] \\
&- \mathcal{R}_i[\Pi_{\bar{\mathcal{M}}_i^\perp}(\Delta_i)].
\end{aligned} \tag{S:8–8}$$

Since $\left\{\widehat{\theta}_i\right\}_{i \in I}$ minimizes the objective function of Eq. (5.2),

$$\begin{aligned}
\sum_{i \in I} \lambda_i \mathcal{R}_i(\widehat{\theta}_i) \le &\sum_{i \in I} \lambda_i \{\mathcal{R}_i(\theta^*_i + \Delta_i) \\
&\mathcal{R}_i[\Pi_{\bar{\mathcal{M}}_i}(\Delta_i)] - \mathcal{R}_i[\Pi_{\bar{\mathcal{M}}_i^\perp}(\Delta_i)]\},
\end{aligned} \tag{S:8–9}$$

Which implies

$$\sum_{i \in I} \lambda_i \mathcal{R}_i[\Pi_{\bar{\mathcal{M}}_i^\perp}(\Delta_i)] \le \sum_{i \in I} \lambda_i \mathcal{R}_i[\Pi_{\bar{\mathcal{M}}_i}(\Delta_i)] \tag{S:8–10}$$

Now, for each structure $i \in I$, we have an application for Hölder's inequality: $|\langle \Delta, \Delta_i \rangle| \le \mathcal{R}^*_i(\Delta)\mathcal{R}_i(\Delta_i) \le 2\lambda_i \mathcal{R}_i(\Delta_i)$ where the notation $\langle\langle A, B \rangle\rangle$ denotes the trace inner product, $\mathrm{trace}(A^T B) = \Sigma_i \Sigma_j A_{ij} B_{ij}$, and we use the pre-computed bound in Eq. (S:8–7). Then, the Frobenius error $||\Delta||_F$ can be upper-bounded as follows:

$$\begin{aligned}
&||\Delta||_F^2 = \langle\langle \Delta, \Delta \rangle\rangle = \sum_{i \in I} \langle\langle \Delta, \Delta_i \rangle\rangle \le \sum_{i \in I} |\langle\langle \Delta, \Delta_i \rangle\rangle| \\
&\le 2 \sum_{i \in I} \lambda_i \mathcal{R}_i(\Delta_i) \le 2 \sum_{i \in I} \{\lambda_i \mathcal{R}_i[\Pi_{\bar{\mathcal{M}}_i}(\Delta_i)] + \\
&\lambda_i \mathcal{R}_i[\Pi_{\bar{\mathcal{M}}_i^\perp}(\Delta_i)]\} \le 4 \sum_{i \in I} \lambda_i \mathcal{R}_i[\Pi_{\bar{\mathcal{M}}_i}(\Delta_i)] \\
&\le 4 \sum_{i \in I} \lambda_i \Psi(\bar{\mathcal{M}}_i) ||\Pi_{\bar{\mathcal{M}}_i}(\Delta_i)||_F
\end{aligned} \tag{S:8–11}$$

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

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

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

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

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

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

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

By this proposition, we have

$$\begin{aligned}&\sum_{i\in I} ||\Delta_i||_F^2 \le ||\Delta||_F^2 + 2|\sum_{i<j}\langle\langle \Delta_i, \Delta_j\rangle\rangle| \\ &\le ||\Delta||_F^2 + \frac{1}{2}\sum_{i\in I}||\Delta_i||_F^2,\end{aligned} \quad \text{(S:8–15)}$$

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

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

$$\begin{aligned}&(\sum_{i\in I} ||\Pi_{\bar{\mathcal{M}}_i}(\Delta_i)||_F)^2 \le (\sum_{i\in I}||\Delta_i||_F)^2 \\ &\le |I|\sum_{i\in I}||\Delta_i||_F^2 \le 8|I|\Phi\sum_{i\in I}||\Pi_{\bar{\mathcal{M}}_i}(\Delta_i)||_F\end{aligned} \quad \text{(S:8–16)}$$

and therefore,

$$\sum_{i\in I} ||\Pi_{\bar{\mathcal{M}}_i}(\Delta_i)||_F \le 8|I|\Phi \quad \text{(S:8–17)}$$

The Frobenius norm error bound Eq. (5.5) can be derived by plugging Eq. (S:8–17) back into Eq. (S:8–13):

$$||\Delta||_F^2 \le 32|I|\Phi^2. \quad \text{(S:8–18)}$$

Therefore, we have

$$||\Delta||_F \le 8\Phi \quad \text{(S:8–19)}$$

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 $i \in I$,

$$\begin{aligned}&\mathcal{R}_i(\Delta_i)\\ &\le \frac{1}{\lambda_i}\{\lambda_i\mathcal{R}_i[\Pi_{\bar{\mathcal{M}}_i}(\Delta_i)] + \lambda_i\mathcal{R}_i[\Pi_{\bar{\mathcal{M}}_i^{\perp}}(\Delta_i)]\}\\ &\le \frac{1}{\lambda_i}\{\lambda_i\mathcal{R}_i[\Pi_{\bar{\mathcal{M}}_i}(\Delta_i)] + \sum_{j\in I}\lambda_j\mathcal{R}_j[\Pi_{\bar{\mathcal{M}}_j}(\Delta_j)]\}\\ &\le \frac{2}{\lambda_i}\sum_{j\in I}\lambda_j\mathcal{R}_j[\Pi_{\bar{\mathcal{M}}_j}(\Delta_j)]\\ &\le \frac{2}{\lambda_i}\sum_{j\in I}\lambda_j\Psi(\bar{\mathcal{M}}_j)||\Pi_{\bar{\mathcal{M}}_j}(\Delta_j)||_F\\ &\le \frac{2\Phi}{\lambda_i}\sum_{j\in I}||\Pi_{\bar{\mathcal{M}}_j}(\Delta_j)||_F \le \frac{16|I|\Phi^2}{\lambda_i} = \frac{32\Phi^2}{\lambda_i}\end{aligned} \quad \text{(S:8–20)}$$

which completes the proof. □

Proof of Theorem (5.4)

Proof. Since $\lambda_n > \lambda'_n$ and $\sqrt{s} > \sqrt{s_{\mathcal{G}}}$, We have that $\lambda_n\sqrt{s} > \lambda'_n\sqrt{s_{\mathcal{G}}}$.
By Theorem (5.3),
$||\widehat{\Omega}_{tot} - \Omega^*_{tot}||_F \le 8\max(\lambda_n\sqrt{s}, \lambda'_n\sqrt{s_{\mathcal{G}}}) \le 8\sqrt{s}\lambda_n$. □

S:8.1 Useful lemma(s)

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

$$|||T_v(\widehat{\Sigma}) - \Sigma|||_\infty \le 5\nu^{1-q}c_0(p) + 3\nu^{-q}c_0(p)\delta \quad \text{(S:8–21)}$$

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

$$||\frac{X^TX}{n} - \Sigma||_\infty \le 8(\max_i \Sigma_{ii})\sqrt{\frac{10\tau\log p'}{n}} \quad \text{(S:8–22)}$$

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

Proof of Corollary (5.5)

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

$$\Omega_{tot} := \begin{pmatrix} \Omega^{(1)} & 0 & \cdots & 0 \\ 0 & \Omega^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Omega^{(K)} \end{pmatrix}$$

The condition (C-SparseΣ) and condition (C-MinInfΣ) also hold for Ω^*_{tot} and Σ^*_{tot}. In order to utilize Theorem (5.4) for this specific case, we only need to show that $||\Omega^*_{tot} - [T_\nu(\widehat{\Sigma}_{tot})]^{-1}||_\infty \leq \lambda_n$ for the setting of λ_n in the statement:

$$\begin{aligned} &||\Omega^*_{tot} - [T_\nu(\widehat{\Sigma}_{tot})]^{-1}||_\infty = ||[T_\nu(\widehat{\Sigma}_{tot})]^{-1}(T_\nu(\widehat{\Sigma}_{tot})\Omega^*_{tot} - I)||_\infty \\ &\leq |||[T_\nu(\widehat{\Sigma}_{tot})w]|||_\infty ||T_\nu(\widehat{\Sigma}_{tot})\Omega^*_{tot} - I||_\infty \\ &= |||[T_\nu(\widehat{\Sigma}_{tot})]^{-1}|||_\infty ||\Omega^*_{tot}(T_\nu(\widehat{\Sigma}_{tot}) - \Sigma^*_{tot})||_\infty \\ &\leq |||[T_\nu(\widehat{\Sigma}_{tot})]^{-1}|||_\infty |||\Omega^*_{tot}|||_\infty ||T_\nu(\widehat{\Sigma}_{tot}) - \Sigma^*_{tot}||_\infty. \end{aligned} \tag{S:8–23}$$

We first compute the upper bound of $|||[T_\nu(\widehat{\Sigma}_{tot})]^{-1}|||_\infty$. By the selection ν in the statement, Lemma (S:8.1) and Lemma (S:8.2) hold with probability at least $1 - 4/p'^{\tau-2}$. Armed with Eq. (S:8–21), we use the triangle inequality of norm and the condition (C-SparseΣ): for any w,

$$\begin{aligned} &||T_\nu(\widehat{\Sigma}_{tot})w||_\infty = ||T_\nu(\widehat{\Sigma}_{tot})w - \Sigma w + \Sigma w||_\infty \\ &\geq ||\Sigma w||_\infty - ||(T_\nu(\widehat{\Sigma}_{tot}) - \Sigma)w||_\infty \\ &\geq \kappa_2 ||w||_\infty - ||(T_\nu(\widehat{\Sigma}_{tot}) - \Sigma)w||_\infty \\ &\geq (\kappa_2 - ||(T_\nu(\widehat{\Sigma}_{tot}) - \Sigma)w||_\infty)||w||_\infty \end{aligned} \tag{S:8–24}$$

Where the second inequality uses the condition (C-SparseΣ). Now, by Lemma (S:8.1) with the selection of ν, we have

$$|||T_\nu(\widehat{\Sigma}_{tot}) - \Sigma|||_\infty \leq c_1 (\frac{\log p'}{n_{tot}})^{(1-q)/2} c_0(p) \tag{S:8–25}$$

where c_1 is a constant related only on τ and $\max_i \Sigma_{ii}$. Specifically, it is defined as $6.5(16(\max_i \Sigma_{ii})\sqrt{10\tau})^{1-q}$. Hence, as long as $n_{tot} > (\frac{2c_1 c_0(p)}{\kappa_2})^{\frac{2}{1-q}} \log p'$ as stated, so that $|||T_\nu(\widehat{\Sigma}_{tot}) - \Sigma|||_\infty \leq \frac{\kappa_2}{2}$, we can conclude that $||T_\nu(\widehat{\Sigma}_{tot})w||_\infty \geq \frac{\kappa_2}{2}||w||_\infty$, which implies $|||[T_\nu(\widehat{\Sigma}_{tot})]^{-1}|||_\infty \leq \frac{2}{\kappa_2}$.

The remaining term in Eq. (S:8–23) is $||T_\nu(\widehat{\Sigma}_{tot}) - \Sigma^*_{tot}||_\infty$; $||T_\nu(\widehat{\Sigma}_{tot}) - \Sigma^*_{tot}||_\infty \leq ||T_\nu(\widehat{\Sigma}_{tot}) - \widehat{\Sigma}_{tot}||_\infty + ||\widehat{\Sigma}_{tot} - \Sigma^*_{tot}||_\infty$. By construction of $T_\nu(\cdot)$ in (C-Thresh) and by Lemma (S:8.2), we can confirm that $||T_\nu(\widehat{\Sigma}_{tot}) - \widehat{\Sigma}_{tot}||_\infty$ as well as $||\widehat{\Sigma}_{tot} - \Sigma^*_{tot}||_\infty$ can be upper-bounded by ν.

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

References

[1] Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

[2] Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep K Ravikumar. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pages 1348–1356, 2009.

[3] Patrick L Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.

[4] Anthony Man-Cho So Zirui Zhou, Qi Zhang. Error bounds and convergence rate analysis of first-order methods. In *Proceedings of the 32th International Conference on Machine learning*, 2015.

[5] Patrick Danaher, Pei Wang, and Daniela M Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2013.

[6] Karl W Broman and Terence P Speed. A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):641–656, 2002.

[7] Elías Moreno, F Javier Girón, and George Casella. Consistency of objective bayes factors as the model dimension grows. *The Annals of Statistics*, pages 1937–1952, 2010.

[8] Yuhong Yang and Andrew R Barron. An asymptotic property of model selection criteria. *Information Theory, IEEE Transactions on*, 44(1):95–116, 1998.

[9] Eunho Yang, Aurélie C Lozano, and Pradeep K Ravikumar. Elementary estimators for graphical models. In *Advances in Neural Information Processing Systems*, pages 2159–2167, 2014.

[10] Dondapati Chowdary, Jessica Lathrop, Joanne Skelton, Kathleen Curtin, Thomas Briggs, Yi Zhang, Jack Yu, Yixin Wang, and Abhijit Mazumder. Prognostic gene expression signatures can be measured in tissues collected in rnalater preservative. *The journal of molecular diagnostics*, 8(1):31–39, 2006.

[11] Michael E Burczynski, Ron L Peterson, Natalie C Twine, Krystyna A Zuberek, Brendan J Brodeur, Lori Casciotti, Vasu Maganti, Padma S Reddy, Andrew Strahs, Fred Immermann, et al. Molecular classification of crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells. *The journal of molecular diagnostics*, 8(1):51–61, 2006.

[12] Erming Tian, Fenghuang Zhan, Ronald Walker, Erik Rasmussen, Yupo Ma, Bart Barlogie, and John D Shaughnessy Jr. The role of the wnt-signaling antagonist dkk1 in the development of osteolytic lesions in multiple myeloma. *New England Journal of Medicine*, 349(26):2483–2494, 2003.

[13] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[14] TS Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, et al. Human protein reference database?2009 update. *Nucleic acids research*, 37(suppl 1):D767–D772, 2009.

[15] Sandra Orchard, Mais Ammari, Bruno Aranda, Lionel Breuza, Leonardo Briganti, Fiona Broackes-Carter, Nancy H Campbell, Gayatri Chavali, Carol Chen, Noemi Del-Toro, et al. The MIntAct project IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic acids research*, page gkt1115, 2013.

[16] Eunho Yang and Pradeep K Ravikumar. Dirty statistical models. In *Advances in Neural Information Processing Systems*, pages 611–619, 2013.

[17] Adam J Rothman, Elizaveta Levina, and Ji Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186, 2009.

[18] Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, Bin Yu, et al. High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.