Appendix to "Learning Latent Variable Gaussian Graphical Models"

A. Motivating Real-World Examples

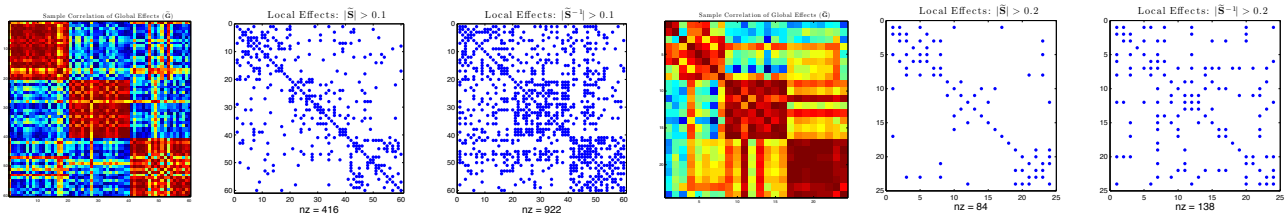
In this section, we use two real-world examples, movie rating and stock return price data sets, to motivate the LVGGM. For each data set, we manually choose three groups of variables where variables in one group are related. Effectively we have injected certain global effects, *i.e.,* group effect, in the data. According to the decomposition of covariance matrix of a LVGGM (see Eq. (20)), we examine whether these effects can be extracted using a low-rank component $\mathbf{G}$ in the covariance matrix, and whether the remaining residual effects have a precision matrix $\mathbf{S}$ that is sparser than its inverse $\mathbf{S}^{-1}$.

We emphasize that for these two examples we are using eigen-decomposition to decompose the covariance matrix into two components. However, this is not related to the regularized ML estimation algorithm proposed in Section 3.4. The low-rank and sparse components that would be learned from the regularized ML problem are different to what we are showing here.

Movielens data. Using the *Movielens*[1] movie rating data set, we choose the rating scores given by the most active 600 users and for the highest rated 20 movies from each of the following three genres: *Horror*, *Children's*, and *Action*. This results in a 600×60 rating matrix with 56% completeness. We consider the joint distribution of 60 movie rating variables as a LVGGM with three latent variables. Each user's rating vector is treated as an *i.i.d.* sample from the LVGGM. Since the true covariance matrix is unknown, we use the sample covariance matrix as a proxy (as $n \gg p$). Each covariance element is weighted by the actual number of observations to compensate for the missingness in the data.

To validate this intuition, we decompose the rating matrix into two matrices: a rank-3 matrix spanned by the top three leading singular vectors, and a residual matrix capturing the conditional effects. We denote the covariance matrix of the low-rank component as $\widetilde{\mathbf{G}}$, and the sparse precision matrix of the residual component as $\widetilde{\mathbf{S}}$. A heat map of the normalized $\widetilde{\mathbf{G}}$ is shown in Figure 4(a), and the sparsity patterns of the normalized $\widetilde{\mathbf{S}}$ and $\widetilde{\mathbf{S}}^{-1}$ (*i.e.,* the covariance of the residual) are shown in Figure 4(b), thresholded by 0.1. As expected, the low-rank $\widetilde{\mathbf{G}}$ captures the structure of the global effects (*i.e.,* genre), and the residual can be well-modeled by a sparse GGM – as we observe that the precision matrix is much sparser than the covariance matrix. In addition, we find the effective rank of the covariance is equal to 7.4, much smaller than the number of variables, 60.

Stock return data. Next, we validate the LVGGM assumptions on a monthly stock return data set[2], which consists of 216 samples of 24 stocks from three sectors: *Technologies*, *Industrials*, and *Financials*. Similar to the *Movielens* data, we reconstruct the low-rank component matrix $\widetilde{\mathbf{G}}$ for the global effects (with rank = 4), and the sparse precision matrix $\widetilde{\mathbf{S}}$ for the residual. The heat map of $\widetilde{\mathbf{G}}$ and the sparsity patterns of $\widetilde{\mathbf{S}}$ and $\widetilde{\mathbf{S}}^{-1}$ are shown in Figure 4(c) and 4(d), respectively. Again, the global structure (*i.e.*, sector) is manifested in the low-rank matrix, and the conditional effects have a much sparser precision matrix than the covariance. We find the effective rank is equal to 2.9, which again is much smaller than the total number of variables, 24.



(a) $\widetilde{\mathbf{G}}$ of *Movielens* (b) $|\widetilde{\mathbf{S}}|$ and $|\widetilde{\mathbf{S}}^{-1}|$ of *Movielens* (c) $\widetilde{\mathbf{G}}$ of stock data (d) $|\widetilde{\mathbf{S}}|$ and $|\widetilde{\mathbf{S}}^{-1}|$ of stock data

Figure 4. Illustration of LVGGM assumptions on *Movielens* and stock return data sets. (a)(c): Heat maps of the leading low-rank matrices capturing the global effects. (b)(d): Sparsity patterns of the precision and covariance matrices of the remaining conditional effects.

[1] http://movielens.org

[2] http://people.csail.mit.edu/myungjin/latentTree.html

B. Proof of Theorem 1

In Yang & Ravikumar (2013), the authors proved a general superpositioned parameter estimate error bound using the decomposable regularized framework. Theorem 1 can be proven similarly by specializing the result in Yang & Ravikumar (2013) to the LVGGM learning problem (4). Then it suffices to verify the two critical conditions (C3) and (C4) in Yang & Ravikumar (2013) (the other two conditions are trivial to verify for our problem), which we introduce and elaborate in this section.

Restricted strong convexity. Let $\delta\mathcal{L}(\Delta;\Theta^*)$ denote the remainder term in first-order Taylor series approximation of the loss function $\mathcal{L}(\cdot)$ at the true parameter Θ^* with respect to a perturbation $\Delta = \Theta^* - \widehat{\Theta}$:

$$\delta\mathcal{L}(\Delta;\Theta^*) := \mathcal{L}(\Theta^* + \Delta) - \mathcal{L}(\Theta^*) - \langle \nabla\mathcal{L}(\Theta^*), \Delta\rangle. \tag{21}$$

In Negahban et al. (2012), the authors introduce the *restricted strong convexity* (RSC) condition, which specifies that given some set $\mathbb{C} \subseteq \mathbb{R}^{p\times p}$, there exists some curvature parameter $\kappa_{\mathcal{L}} > 0$ and tolerance function $\tau_{\mathcal{L}}$, such that the following holds:

$$\delta\mathcal{L}(\Delta;\Theta^*) \geq \kappa_{\mathcal{L}}\|\Delta\|_F^2 - \tau_{\mathcal{L}}(\Theta^*),\ \forall\Delta \in \mathbb{C}. \tag{22}$$

The RSC condition guarantees sufficient curvature of the loss function at the true parameter along some directions specified by set $\mathbb{C}$. This condition is critical for consistent estimation in the high-dimensional regime, since standard strong convexity usually does not hold in the $p \gg n$ setting.

The following shows that the restricted Fisher eigenvalue conditions defined in Assumption 1 implies the RSC condition.

Lemma 2 (RSC condition)**.** *Suppose Assumption 1 holds for the true marginal precision matrix Θ^* and let $M > 2$. Then for all $\Delta \in \mathbb{C}(E) \cup \mathbb{C}(U)$, such that $\|\Delta\|_{\mathcal{F}^*}^2 \leq \frac{1}{2M^2}$, the RSC condition (22) is satisfied with the curvature parameter $\kappa_{\mathcal{L}} = \frac{M-2}{2(M-1)}\kappa^*_{\min}$ and the tolerance function $\tau_{\mathcal{L}} = 0$.*

The proof of Lemma 2 is largely inspired by Kakade et al. (2010), in which it is shown that exponential family distributions exhibit *almost strong convexity* in a neighborhood. The RFE assumption makes connection between this property and the RSC condition. A proof of Lemma 2 is given in the Appendix C.

Note there is an important difference between the RSC condition considered here and the condition introduced in Agarwal et al. (2012). The RSC condition considered here is satisfied with respect to the error matrices of each simple structure separately, while the RSC condition in Agarwal et al. (2012) is required for the combined error matrices (defined in the product space of two sets), which could lie in a significantly larger set.

Structural incoherence. The second ingredient for consistent estimation of the sparse plus low-rank parameter Θ, is some type of incoherence condition between the sparse and low-rank components. In the present work, we consider the *structural incoherence* condition that was proposed more recently in (Yang & Ravikumar, 2013). This condition allows for a vanishing error bound when n goes to infinity, and is applicable to more general loss functions, such as the log-likelihood function in Eq. (3).

Define the following incoherence measure of the loss function $\mathcal{L}$ for two structural error matrices Δ_S and Δ_L:

$$\begin{aligned} c_{\mathcal{L}}(\Delta_S, \Delta_L; \Theta^*) := |&\mathcal{L}(\Theta^* + \Delta_S + \Delta_L) + \mathcal{L}(\Theta^*) \\ &- \mathcal{L}(\Theta^* + \Delta_S) - \mathcal{L}(\Theta^* + \Delta_L)|, \forall \Delta_S \in \mathbb{C}(E), \Delta_L \in \mathbb{C}(U). \end{aligned}$$

Then the *structural incoherence* (SI) condition is satisfied if the following relation holds for all $\Delta_S \in \mathbb{C}(E)$ and $\Delta_L \in \mathbb{C}(U)$:

$$c_{\mathcal{L}}(\Delta_S, \Delta_L; \Theta^*) \leq \frac{\kappa_{\mathcal{L}}}{2}\|(\|\Delta_S\|_F^2 + \|\Delta_L\|_F^2), \tag{23}$$

where $\kappa_{\mathcal{L}}$ is the curvature parameter in the RSC condition (22).

The following lemma shows that, in addition to the restricted Fisher eigenvalue assumption (Assumption 1), if the true marginal model also satisfies the structural Fisher incoherence assumption (Assumption 2), then the above SI condition on the likelihood loss function is guaranteed.

Lemma 3 (SI condition). *Suppose Assumption 1 and 2 hold for the true marginal precision matrix Θ^* and let $M > 6$. Then the SI condition (23) is satisfied for all $\Delta_S \in \mathbb{C}(E)$ and $\Delta_L \in \mathbb{C}(U)$, such that $\max\{\|\Delta_S\|_{\mathcal{F}^*}^2, \|\Delta_L\|_{\mathcal{F}^*}^2\} \leq \frac{1}{6M^2}$. The curvature parameter $\kappa_{\mathcal{L}}$ is the same as in Lemma 2, i.e., $\kappa_{\mathcal{L}} = \frac{M-2}{2(M-1)}\kappa_{\min}^*$.*

The proof of Lemma 3 is in Section D in Appendix.

Finally, under Assumption 1 and Assumption 2, Lemma 2 and 3 imply the RSC and SI conditions hold for our LVGGM learning problem, respectively. Thus Theorem 1 can be proven by directly appealing to Theorem 1 in (Yang & Ravikumar, 2013).

C. Proof of Lemma 2

Proof. The remainder term in the first-order Taylor series of the negative log-likelihood (3) of GGM takes the following form:

$$\begin{aligned}\delta\mathcal{L}(\Delta;\Theta^*) &= \mathcal{L}(\Theta^*+\Delta) - \mathcal{L}(\Theta^*) - \langle\nabla\mathcal{L}(\Theta^*),\Delta\rangle \\ &= \langle{\Theta^*}^{-1},\Delta\rangle - \log\det(\Theta^*+\Delta) + \log\det(\Theta^*).\end{aligned}$$

For $s \in (0,1]$, define the Taylor series of function $g(s;\Theta^*) := \log\det(\Theta^* + s\Delta)$ at Θ^*

$$g(s;\Theta^*) = \log\det(\Theta^*+s\Delta) = \sum_{k=0}^{\infty}\frac{c_k(\Delta)s^k}{k!}, \tag{24}$$

where $c_k(\Delta) := g^{(k)}(s;\Theta^*)$ is the k-th derivative of the $\log\det$ function at Θ^*. Define $c_0(\Delta) := \log\det(\Theta^*)$, the remainder can be expressed as:

$$\delta\mathcal{L}(s\Delta;\Theta^*) = \sum_{k=2}^{\infty}\frac{c_k(\Delta)s^k}{k!} = \frac{c_2(\Delta)s^2}{2} + \sum_{k=3}^{\infty}\frac{c_k(\Delta)s^k}{k!} = \frac{c_2(\Delta)s^2}{2} + \delta g(s;\Delta,\Theta^*), \tag{25}$$

where the second term $\delta g(s)$ is defined as the second-order Taylor error of the log-determinant function. Next we show that this error term, which is the sum of all the higher-order terms, can be bounded by a quadratic term in a small neighborhood around Θ^*.

For exponential family distributions (Gaussian as an example), the log-partition function (*i.e.,* $\log\det$ function for Gaussian) coincides with the *cumulant generating function*. This implies that the derivatives $c_k(\Delta)$ are the corresponding cumulants of the distribution, which can be shown to converge to zero quite rapidly. Indeed, in Kakade et al. (2010) the authors show that for a univariate random variable z under an exponential family distribution, its k-th order cumulant satisfies

$$\left|\frac{c_k(z)}{c_2(z)^{k/2}}\right| \leq \frac{1}{2}k!\alpha^{k-2}, \quad \forall k \geq 3, \tag{26}$$

where α is a finite constant, and the second-order cumulant coincides with the Fisher norm of the deviation $c_2(\Delta) = \|\Delta\|_{\mathcal{F}^*}^2$ due to the definition of the Fisher information. For multivariate Gaussian distributions, $\alpha = \sqrt{2}$ suffices for the above relation to hold (see Sec. 3.2.2 in Kakade et al. (2010)).

Therefore we bound the second-order Taylor error term in Eq. (25) as follows (similar to Kakade et al. (2010)):

$$|\delta g(s;\Delta,\Theta^*)| = \left|\sum_{k=3}^{\infty}\frac{c_k(\Delta)s^k}{k!}\right| \tag{27}$$

$$\leq \frac{1}{2}\sum_{k=3}^{\infty}2^{\frac{k}{2}-1}c_2(\Delta)^{k/2}s^k \tag{28}$$

$$\leq \frac{s^2c_2(\Delta)}{2}\sum_{k=1}^{\infty}(s\sqrt{2c_2(\Delta)})^k \tag{29}$$

$$\overset{(i)}{\leq} \frac{s^2c_2(\Delta)}{2}\sum_{k=1}^{\infty}\frac{1}{M^k} \tag{30}$$

$$= \frac{s^2c_2(\Delta)}{2(M-1)} \tag{31}$$

$$\leq \frac{c_2(\Delta)}{2(M-1)}\frac{1}{\max\{2M^2c_2(\Delta),1\}} \tag{32}$$

$$\overset{(ii)}{=} \frac{c_2(\Delta)}{2(M-1)} \tag{33}$$

where (i) and (ii) are due to our conditions on $c_2(\Delta)$ (*i.e.,* $\|\Delta\|_{\mathcal{F}^*}^2 \leq \frac{1}{2M^2}$) and $s \leq 1$. Then we obtain a lower bound for $\delta\mathcal{L}(\Delta;\Theta^*)$:

$$\delta\mathcal{L}(\Delta;\Theta^*) \geq \frac{c_2(\Delta)}{2} + \delta g(s;\Delta,\Theta^*) \geq \left(\frac{1}{2}-\frac{1}{2(M-1)}\right)c_2(\Delta) \overset{(ii)}{\geq} \frac{M-2}{2(M-1)}\kappa_{\min}^*\|\Delta\|_F^2, \tag{34}$$

where (ii) is due to the RFE condition. Therefore the RSC condition is satisfied with the curvature parameter $\kappa_{\mathcal{L}} := \frac{M-2}{2(M-1)}\kappa_{\min}^*$ and a zero tolerance parameter $\tau_{\mathcal{L}} = 0$.

□

D. Proof of Lemma 3

Proof. First we state the following lemma which gives a bound on the magnitude of *Fisher inner product* between elements from the two sets.

Lemma 4. *Suppose Assumption 1 and 2 hold for the true marginal precision matrix Θ^*. Then given a constant $M \geq 6$, the following inequality holds for all $\Delta_S \in \mathbb{C}(E)$ and $\Delta_L \in \mathbb{C}(U)$ such that $\max\{\|\Delta_S\|_{\mathcal{F}^*}^2, \|\Delta_L\|_{\mathcal{F}^*}^2\} \leq \frac{1}{6M^2}$:*

$$|\langle\Delta_S,\Delta_L\rangle_{\mathcal{F}^*}| \leq \psi\left(\|\Delta_S\|_{\mathcal{F}^*}^2 + \|\Delta_L\|_{\mathcal{F}^*}^2\right), \tag{35}$$

where $\psi := \frac{1}{4} - \frac{3}{2M}$.

The proof of Lemma 4 follows similarly as that of the Proposition 2 in Yang & Ravikumar (2013), and hence is omitted.

Next we prove Lemma 3 using the above result. Following similar derivations as in the proof of Lemma 2, the incoherence measure in the SI condition can be simplified to

$$c_{\mathcal{L}}(\Delta_S,\Delta_L;\Theta^*) := |\delta\mathcal{L}(\Delta_S+\Delta_L;\Theta^*) - \delta\mathcal{L}(\Delta_S;\Theta^*) - \delta\mathcal{L}(\Delta_L;\Theta^*)|,$$

Using the remainder in the Taylor series of $\delta\mathcal{L}$ (25), the incoherence measure can be expressed as:

$$
\begin{aligned}
&c_{\mathcal{L}}(\Delta_S, \Delta_L; \Theta^*) \\
=&\left| \frac{c_2(\Delta_S + \Delta_L)}{2} + \delta g(s; \Delta_S + \Delta_L) - \left(\frac{c_2(\Delta_S)}{2} + \delta g(s_1; \Delta_S) \right) - \left(\frac{c_2(\Delta_L)}{2} + \delta g(s_2; \Delta_L) \right) \right| \\
\leq& \left| \frac{c_2(\Delta_S + \Delta_L)}{2} - \frac{c_2(\Delta_S)}{2} - \frac{c_2(\Delta_L)}{2} \right| + |\delta g(s; \Delta_S + \Delta_L)| + |\delta g(s_1; \Delta_S)| + |\delta g(s_2; \Delta_L)| \\
\overset{(i)}{\leq}& |\langle \Delta_S, \Delta_L \rangle_{\mathcal{F}^*}| + \frac{c_2(\Delta_S + \Delta_L) + c_2(\Delta_S) + c_2(\Delta_L)}{2(M-1)} \\
=& |\langle \Delta_S, \Delta_L \rangle_{\mathcal{F}^*}| + \frac{\|\Delta_S\|_{\mathcal{F}^*}^2 + \|\Delta_L\|_{\mathcal{F}^*}^2 + \langle \Delta_S, \Delta_L \rangle_{\mathcal{F}^*}}{M-1} \\
\leq& \frac{M}{M-1} |\langle \Delta_S, \Delta_L \rangle_{\mathcal{F}^*}| + \frac{\|\Delta_S\|_{\mathcal{F}^*}^2 + \|\Delta_L\|_{\mathcal{F}^*}^2}{M-1} \\
\overset{(ii)}{\leq}& \frac{M\psi + 1}{M-1} (\|\Delta_S\|_{\mathcal{F}^*}^2 + \|\Delta_L\|_{\mathcal{F}^*}^2) \\
\overset{(iii)}{\leq}& \frac{M-2}{4(M-1)} \kappa_{\min}^* (\|\Delta_S\|_F^2 + \|\Delta_L\|_F^2) \\
\leq& \frac{\kappa_{\mathcal{L}}}{2} (\|\Delta_S\|_F^2 + \|\Delta_L\|_F^2),
\end{aligned}
$$

where in (i) we have apply (33) to bound the second-order Taylor error terms (note that the conditions on the error matrices also guarantees $\|\Delta_S + \Delta_L\|_{\mathcal{F}^*}^2 \leq \frac{1}{2M^2}$ due to Lemma 4). Inequality (ii) is due to Lemma 4. Inequality (iii) can be verified by the definitions of ψ and the RSC curvature parameter $\kappa_{\mathcal{L}}$.

□

E. Proof of Corollary 1

Proof. Theorem 1 is a deterministic statement, however, the condition on the regularization parameters (11) and the error bound depend on the sample covariance matrix $\widehat{\Sigma}$ which is random. Note that the error bound directly follows from the deterministic error bound in Theorem 1 and the choices of regularization parameters as in Eq. (16). To prove Corollary 1, it only remains to verify that the condition (11) in Theorem 1 is guaranteed with high probability. More specifically, this requires bounding the deviation of the sample covariance matrix in terms of ℓ_∞ and and spectral norms.

First we make use of the following lemma to characterize the element-wise deviation of the sample covariance matrix[3].

Lemma 5 (Ravikumar et al. (2011)). *For a p-dimensional Gaussian random vector with covariance matrix $\mathbf{\Sigma}^*$, the sample covariance matrix obtained from n samples $\widehat{\mathbf{\Sigma}}$ satisfies*

$$
P\left\{ |\widehat{\mathbf{\Sigma}}_{i,j} - \mathbf{\Sigma}^*_{i,j}| > \epsilon_1 \right\} \leq 4 \exp\left(-\frac{n\epsilon_1^2}{3200 \overline{\sigma}^{*2}} \right), \tag{36}
$$

for all $\epsilon_1 \in (0, 40\overline{\sigma})$, where $\overline{\sigma}^* := \max_{i=1,\ldots,p} \mathbf{\Sigma}^*_{i,i}$.

If the number of samples satisfies $n \geq 4 \log p$, then by choosing $\frac{1}{2}\lambda \geq \epsilon_1 = 80 C_1 \overline{\sigma}^* \sqrt{\frac{\log p^2}{n}} \in (0, 40\overline{\sigma})$, where $C_1 > 1$ is an arbitrary constant, and applying the union bound we have

$$
P\left\{ \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_\infty \leq \frac{1}{2}\lambda \right\} \geq P\left\{ \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_\infty \leq \epsilon_1 \right\} \geq 1 - 4p^{-2(C_1 - 1)}.
$$

Then the condition on λ is satisfied with high probability.

Next we consider the condition on the other regularization parameter μ, which requires bounding the deviation of the operation norm of the sample covariance matrix. The following lemma provides such a characterization.

[3]The original lemma applies to all sub-Gaussian variables, here we specialize to Gaussian random vectors.

Lemma 6 (Chandrasekaran et al. (2012), Lemma 3.9)**.** *For a p-dimension Gaussian random vector with covariance matrix $\mathbf{\Sigma}^*$ and let $\rho^* = \|\mathbf{\Sigma}^*\|_2$. If the number of samples n be such that $n \geq \frac{64p\rho^{*2}}{\epsilon_2^2}$, then the sample covariance matrix $\widehat{\mathbf{\Sigma}}$ obtained from n samples satisfies*

$$P\left\{\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_2 \geq \epsilon_2\right\} \leq 2\exp\left(-\frac{n\epsilon_2^2}{128\rho^{*2}}\right), \tag{37}$$

for all $\epsilon_2 \in (0, 8\rho^*)$.

If $n \geq p$, then by choosing $\frac{1}{2}\mu \geq \epsilon_2 = 8C_2\rho^*\sqrt{\frac{p}{n}} \in (0, 8\rho^*)$, where $C_2 \geq 1$ is an arbitrary constant, we have

$$P\left\{\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_2 \leq \frac{1}{2}\mu\right\} \geq P\left\{\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_2 \leq \epsilon_2\right\} \geq 1 - 2\exp\left(-\frac{C_2^2 p}{2}\right).$$

Combining the above results we have verified the condition (11) in Theorem 1 holds with high probability, which concludes the proof. □