Supplementary Materials for “HMLasso: Lasso with High Missing Rate”

A Discussions from Asymptotic Perspective

We describe another view of our weighted formulation. Note that this is a rough result, but we intuitively interpret why weighted norm with $\alpha = 1/2$ performs well from the asymptotic perspective ($n \gg p$).

The standard asymptotic theory shows that we have, for a large pairwise observation number n_{jk},

$$\sqrt{n_{jk}}(S_{jk}^{\mathrm{pair}} - \Sigma_{jk}^*) \sim \mathcal{N}(0, \tau_{jk}^2),$$

where Σ^* is a population covariance matrix and τ_{jk} is a constant. Here we assume that S_{jk}^{pair}’s are independent and $\tau_{jk} = \tau$ for all j, k. Then, the likelihood of S^{pair} can be approximated to

$$\prod_{j,k} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2\tau^2}\left(\sqrt{n_{jk}}(S_{jk}^{\mathrm{pair}} - \Sigma_{jk}^*)\right)^2\right).$$

Hence, the maximum likelihood estimator of Σ^* under the PSD constraint can be approximated to

$$\operatorname*{argmin}_{\Sigma \succeq 0} \sum_{j,k} n_{jk}(S_{jk}^{\mathrm{pair}} - \Sigma_{jk})^2,$$

which is equivalent to our method with $\alpha = 1/2$.

B Algorithms

B.1 Cordinate Descent Algorithm with the Covariance Matrix

We describe an algorithm for solving the Lasso problem using the covariance matrix (8). Let $\mathscr{L}(\beta)$ be the objective function of (8). To derive the update equation, when $\beta_j \neq 0$, differentiating $\mathscr{L}(\beta)$ with respect to β_j yields

$$\partial_{\beta_j}\mathscr{L}(\beta) = \tilde{\Sigma}_{j,-j}\beta_{-j} + \tilde{\Sigma}_{jj}\beta_j - \rho_j^{\mathrm{pair}} + \lambda \operatorname{sgn}(\beta_j),$$

Algorithm 1 Lasso with Covariance Matrix

```
Input: $\tilde{\Sigma}$, $\rho^{\mathrm{pair}}$, $\lambda$
  initialize $\beta$
  while until convergence do
      for $j = 1, \cdots, p$ do
          $\beta_j \leftarrow \frac{1}{\tilde{\Sigma}_{jj}} S\left(\left(\rho_j^{\mathrm{pair}} - \tilde{\Sigma}_{j,-j}\beta_{-j}\right), \lambda\right)$
      end for
  end while
Output: $\beta$
```

where β_{-j} denotes β without the j-th component, and $X_{j,-j}$ denotes the j-th row of X without the j-th column. Solving $\partial_{\beta_j}\mathscr{L}(\beta) = 0$, we obtain the update rule as

$$\beta_j \leftarrow \frac{1}{\tilde{\Sigma}_{jj}} S\left(\left(\rho_j^{\mathrm{pair}} - \tilde{\Sigma}_{j,-j}\beta_{-j}\right), \lambda\right),$$

where $S(z, \gamma)$ is a soft thresholding function

$$\begin{aligned} S(z,\gamma) :&= \mathrm{sgn}(z)(|z| - \gamma)_+ \\ &= \begin{cases} z - \gamma & \text{if } z > 0 \text{ and } \gamma < |z|, \\ z + \gamma & \text{if } z < 0 \text{ and } \gamma < |z|, \\ 0 & \text{if } |z| \leq \gamma. \end{cases} \end{aligned}$$

The whole algorithm for the Lasso-type optimization problem (8) is described in Algorithm 1.

B.2 ADMM for the Weighted Max Norm Formulation

We describe the ADMM algorithm for the weighted max norm formulation. This is a natural extension of the CoCoLasso algorithm. The difference from the CoCoLasso algorithm is B-step update in ADMM described in Algorithm 2.

C Proofs

C.1 Proof of Proposition 1

Proof. First, we prove that random variables with Bernoulli distribution is sub-Gaussian. Because $m_{ij} \sim Bernoulli(\mu_j)$, we have $E[m_{ij}] = \mu_j$ and

$$\begin{aligned} &E[\exp(s(m_{ij} - \mu_j))] \\ =&\mu_j \exp(s(1 - \mu_j)) + (1 - \mu_j)\exp(-s\mu_j). \end{aligned}$$

Algorithm 2 B-step Update for max norm in ADMM

```
**Input:** $A_{k+1}, \Lambda_k, \hat{\Sigma}, \mu, W$
  define $c = \mathrm{vec}\left(A_{k+1} - \hat{\Sigma} - \mu\Lambda_k\right)$, $w = \mathrm{vec}(W)$
  sort $c$ as $w_1|c_1| \geq w_2|c_2| \geq \dots$
  find $l = \max_{l'} \left\{ l' : w_l'|c_l'| - \frac{\left(\sum_{j=1}^{l'} |c_j|\right) - \frac{\mu}{2}}{\sum_{j=1}^{l'} \frac{1}{w_j}} > 0 \right\}$
  define $d = \frac{\left(\sum_{j=1}^{l} |c_j|\right) - \frac{\mu}{2}}{\sum_{j=1}^{l} \frac{1}{w_j}}$
  define $B_{k+1} = \mathrm{mat}(b)$ such that $b_j = c_j$ for $|c_j| \leq \frac{d}{w_j}$, and $b_j = \frac{d\mathrm{sgn}(c_j)}{w_j}$ for $|c_j| > \frac{d}{w_j}$
**Output:** $B_{k+1}$
```

Hence, the Taylor expansion yields

$$\begin{aligned}
&\log E[\exp(s(m_{ij} - \mu_j))] \\
=& - s\mu_j + \log(1 + \mu_j(\exp(s) - 1)) \\
\leq& \frac{\mu_j(1-\mu_j)s^2}{2},
\end{aligned}$$

which indicates m_{ij} is sub-Gaussian with $\tau_j^2 = \mu_j(1-\mu_j)$. We can see $\tau_j^2 \leq 1/4$ since $\mu_j \in [0,1]$.

Next, we prove the proposition. For a random vector M_i, the i-th row of M, we have

$$\begin{aligned}
&E[\exp(s(v^\top M_i - E[v^\top M_i]))] \\
=&E[\exp(s(v^\top M_i - E[v^\top \mu]))] \\
=&\prod_{j=1}^{p} E[\exp(sv_j(M_{ij} - \mu_j))] \\
\leq&\prod_{j=1}^{p} \exp\left(\frac{\tau_j^2 v_j^2 s^2}{2}\right) \\
=&\exp\left(\frac{\left(\sum_{j=1}^{p} \tau_j^2 v_j^2\right) s^2}{2}\right) \\
\leq&\exp\left(\frac{s^2 \max_j \tau_j^2}{2}\right),
\end{aligned}$$

for any unit vector v.

□

C.2 Proof of Theorem 2

Proof. We see that

$$\begin{aligned}\left|\hat{\Sigma}_{jk}-S_{jk}\right| &= \left|\frac{1}{n}\sum_{i=1}^{n} m_{ij}m_{ik}x_{ij}x_{ik}/r_{jk}-\frac{1}{n}\sum_{i=1}^{n}x_{ij}x_{ik}\right|\\ &\leq \frac{1}{r_{jk}}\frac{1}{n}\left|\sum_{i=1}^{n}x_{ij}x_{ik}(m_{ij}-\mu_j)(m_{ik}-\mu_k)\right|\\ &+\frac{\mu_j}{r_{jk}}\frac{1}{n}\left|\sum_{i=1}^{n}x_{ij}x_{ik}(m_{ik}-\mu_k)\right|\\ &+\frac{\mu_k}{r_{jk}}\frac{1}{n}\left|\sum_{i=1}^{n}x_{ij}x_{ik}(m_{ij}-\mu_j)\right|.\end{aligned}$$

We denote the three terms on the right-hand side by T_1, T_2, and T_3, respectively.

(T1): Let $v_i := x_{ij}x_{ik}$. Then we have $\|v\|_\infty \leq X_{\max}^2$. Remember that $m_{ij}-\mu_j$ and $m_{ik}-\mu_k$ are sub-Gaussian with parameter τ^2. Then, by applying Lemma B.1 in CoCoLasso, we have

$$\begin{aligned}&\Pr(T_1>\varepsilon)\\ =&\Pr\left(\frac{1}{n}\left|\sum_{i=1}^{n}x_{ij}x_{ik}(m_{ij}-\mu_j)(m_{ik}-\mu_k)\right|>r_{jk}\varepsilon\right)\\ \leq& C\exp\left(-\frac{cn\varepsilon r_{jk}^2}{\tau^4X_{\max}^4}\right)\end{aligned}$$

for all $r_{jk}\varepsilon \leq c\tau^2X_{\max}^2$, i.e., $\varepsilon \leq c\tau^2X_{\max}^2/r_{jk}$.

(T2) and (T3): By property (B.2) in CoCoLasso, we can see that for any vector v and independent sub-Gaussian vector w_i with parameter τ^2, we have

$$\Pr\left(\frac{1}{n}\left|\sum_{i=1}^{n}v_iw_i\right|>\varepsilon\right)\leq C\exp\left(-\frac{cn^2\varepsilon^2}{\|v\|_2^2\tau^2}\right).$$

If we define $v_i := x_{ij}x_{ik}$, we have $\|v\|_2^2 \leq nX_{\max}^4$. Remember that $m_{ij}-\mu_j$ and $m_{ik}-\mu_k$ are sub-Gaussian with parameter τ^2. Hence, we have

$$\begin{aligned}&\Pr(T_2>\varepsilon)\\ =&\Pr\left(\frac{1}{n}\left|\sum_{i=1}^{n}x_{ij}x_{ik}(m_{ik}-\mu_k)\right|>\frac{r_{jk}\varepsilon}{\mu_j}\right)\\ \leq& C\exp\left(-\frac{cn\varepsilon^2r_{jk}^2}{\tau^2X_{\max}^4\mu_j^2}\right).\end{aligned}$$

Similarly, we have

$$\begin{aligned}&\Pr(T_3 > \varepsilon)\\=&\Pr\left(\frac{1}{n}\left|\sum_{i=1}^{n} x_{ij}x_{ik}(m_{ij}-\mu_j)\right| > \frac{r_{jk}\varepsilon}{\mu_k}\right)\\\le& C\exp\left(-\frac{cn\varepsilon^2 r_{jk}^2}{\tau^2 X_{\max}^4 \mu_k^2}\right).\end{aligned}$$

Putting all together, we obtain that for all $\varepsilon \le c\tau^2 X_{\max}^2 / r_{jk}$,

$$\begin{aligned}&\Pr\left(\left|\hat{\Sigma}_{jk}-S_{jk}\right| > \varepsilon\right)\\\le& C\exp\left(-\frac{cn\varepsilon^2 r_{jk}^2}{\tau^2 X_{\max}^4 \max\{\tau^2,\mu_j^2,\mu_k^2\}}\right).\end{aligned}$$

□

C.3 Proof of Theorem 3

Proof. Since $\tilde{\Sigma} = \underset{\Sigma\succeq 0}{\operatorname{argmin}} \|W \odot (\Sigma - \hat{\Sigma})\|_{\mathrm{F}}^2$, we have, using the triangular equation,

$$\begin{aligned}&\|W\odot(\tilde{\Sigma}-S)\|_{\mathrm{F}}\\\le&\|W\odot(\tilde{\Sigma}-\hat{\Sigma})\|_{\mathrm{F}}+\|W\odot(\hat{\Sigma}-S)\|_{\mathrm{F}}\\\le&2\|W\odot(\hat{\Sigma}-S)\|_{\mathrm{F}}.\end{aligned}$$

From Theorem 2, we have

$$\begin{aligned}&\Pr\left(\|W\odot(\hat{\Sigma}-S)\|_{\mathrm{F}} > \varepsilon\right)\\=&\Pr\left(\sum_{j,k} W_{jk}^2\left(\hat{\Sigma}_{jk}-S_{jk}\right)^2 > \varepsilon^2\right)\\\le&\sum_{j,k}\Pr\left(W_{jk}^2\left(\hat{\Sigma}_{jk}-S_{jk}\right)^2 > \varepsilon^2 p^{-2}\right)\\\le& p^2\max_{j,k}\Pr\left(W_{jk}\left|\hat{\Sigma}_{jk}-S_{jk}\right| > \varepsilon p^{-1}\right)\\\le& p^2 C\exp\left(-cn\varepsilon^2 p^{-2}\left(\min_{j,k}\frac{r_{jk}}{w_{jk}}\right)^2\zeta^{-1}\right),\end{aligned}$$

for all $\varepsilon \leq cp\tau^2 X_{\max}^2 \min_{j,k} (w_{jk}/r_{jk})$, where $\zeta = \max\left\{\tau^2, \mu_1^2, \ldots, \mu_p^2\right\}$. Hence, we have

$$\begin{aligned}
&\Pr\left(\|W \odot (\tilde{\Sigma} - S)\|_{\mathrm{F}} > \varepsilon\right)\\
\leq &\Pr\left(\|W \odot (\hat{\Sigma} - S)\|_{\mathrm{F}} > \varepsilon/2\right)\\
\leq &p^2 C \exp\left(-cn\varepsilon^2 p^{-2} \left(\min_{j,k} \frac{r_{jk}}{w_{jk}}\right)^2 \zeta^{-1}\right),
\end{aligned}$$

for all $\varepsilon \leq cp\tau^2 X_{\max}^2 \min_{j,k} \left(\frac{w_{jk}}{r_{jk}}\right)$. This is equivalent to

$$\begin{aligned}
&\Pr\left(\frac{1}{p^2} \left\|W \odot (\tilde{\Sigma} - S)\right\|_{\mathrm{F}}^2 > \varepsilon^2\right)\\
\leq &p^2 C \exp\left(-cn\varepsilon^2 \left(\min_{j,k} \frac{r_{jk}}{W_{jk}}\right)^2 \zeta^{-1}\right).
\end{aligned}$$

Using the inequality $W_{\min}^2 \left\|\tilde{\Sigma} - S\right\|_{\mathrm{F}}^2 \leq \left\|W \odot \left(\tilde{\Sigma} - S\right)\right\|_{\mathrm{F}}^2$, we have

$$\begin{aligned}
&\Pr\left(\frac{1}{p^2} \left\|\tilde{\Sigma} - S\right\|_{\mathrm{F}}^2 > \varepsilon^2\right)\\
\leq &p^2 C \exp\left(-cn\varepsilon^2 W_{\min}^2 \left(\min_{j,k} \frac{r_{jk}}{W_{jk}}\right)^2 \zeta^{-1}\right),
\end{aligned}$$

for $\varepsilon \leq c\tau^2 X_{\max}^2 (\min_{j,k} W_{jk}/r_{jk})/W_{\min}$. □

D Numerical Experiments

We discribe additional simulation results. The conditions include various missing patterns, covariance patterns, noise levels, and true parameters.

D.1 Missing Patterns

We examined three missing patterns and three missing rates, resulting in nine conditions. We set the missing rates to $\mu = 0.1, 0.5, 0.9$. We introduced missing values according to the following missing patterns, which are thought to be common in real-world data.

(1) Random pattern: Missing elements were selected with the same probability for all the elements.

(2) Column pattern: Missing rates differ for each column. The j-th column missing rate μ_j was sampled from the uniform distribution so that the overall missing rate was μ. μ_j was sampled from $U(0, 0.2)$ for $\mu = 0.1$, from $U(0, 1)$ for $\mu = 0.5$, and from $U(0.8, 1)$ for $\mu = 0.9$.

(3) Row column pattern: Missing rates differ for each row and each column. The (i, j)-th element missing rate μ_{ij} was set so that the overall missing rate was μ. Specifically, we defined $\mu_{ij} = \mu^i \mu_j$ where μ^i and μ_j were sampled from $U(0, 0.632)$ for $\mu = 0.1$, $\mu_{ij} = 1 - (1 - \mu^i)(1 - \mu_j)$ where μ^i and μ_j were sampled from $U(0.586, 1)$ for $\mu = 0.5$, and $\mu_{ij} = 1 - (1 - \mu^i)(1 - \mu_j)$ where μ^i and μ_j were sampled from $U(0.368, 1)$ for $\mu = 0.9$.

Figure 1 shows the results. HMLasso outperformed other methods, when the missing rate was moderate or high. In particular, in the cases of the column pattern and row column pattern, HMLasso delivered significant improvements. This might be because the number of pairwise observations were very small for these missing patterns. The mean imputation and CoCoLasso suffered from highly missing variables, while HMLasso suppressed the effects of them.

Note that the column and row missing patterns often appear in practice. The column missing pattern appears when some variables are frequently observed and others are rarely observed. This is typically caused by different data collection cost for each variable. The row missing pattern appears when some samples are filled and other samples are highly missing. This happens when some samples are considered to be important and they are frequently measured.

D.2 Covariance Patterns

We examined three covariance patterns and three covariance levels, resulting in nine conditions. We set the covariance levels to $r = 0.1, 0.5, 0.9$. The covariance matrix was generated according to the following covariance matrix patterns.

(1) Uniform pattern: Covariances were uniform among all variables, where $\Sigma^*_{jk} = r$ for $j \neq k$ and $\Sigma^*_{jk} = 1$ for $j = k$.

(2) Autoregressive pattern: Covariances among neighbors were strong, such that $\Sigma^*_{jk} = r^{|j-k|}$ for $j \neq k$ and $\Sigma^*_{jk} = 1$ for $j = k$.

(3) Block pattern: All of the variables were divided into some blocks. The intra-block covariances were strong and inter-block covariances are zeros. We set $\Sigma^* = \mathrm{diag}(\Sigma^*_{11}, \dots, \Sigma^*_{qq})$ with $q = 10$, where Σ^*_{jj} was a 10-dimensional square matrix with the above uniform pattern.

Figure 2 shows the results. HMLasso outperformed the other methods for almost all covariance patterns and covariance levels. The mean imputation method was comparable to HMLasso under low covariance conditions, because the shrinkage estimator such as the mean imputation tends to show a good performance when the true covariance is close to zero. However, the mean imputation deteriorated its estimation under a moderate or high covariance condition.

D.3 Noise Levels and True Parameters

We examined three noise levels and three kinds of true parameters. We set noise levels to $\mathrm{Var}[\varepsilon] = 0.1^2, 1, 10^2$. The true parameters β were defined as the following.

(1) $\beta_1 = 10, \beta_{11} = -9, \beta_{21} = 8, \beta_{31} = -7, \ldots, \beta_{91} = -1$, and $\beta_j = 0$ otherwise.

(2) $\beta_1 = 10, \beta_2 = -9, \beta_3 = 8, \beta_4 = -7, \ldots, \beta_{10} = -1$, and $\beta_j = 0$ otherwise.

(3) $\beta_1 = 5, \beta_{11} = -5, \beta_{21} = 5, \beta_{31} = -5, \ldots, \beta_{91} = -5$, and $\beta_j = 0$ otherwise.

Figure 3 shows the results. HMLasso outperformed the other methods for all noise levels and true parameters.

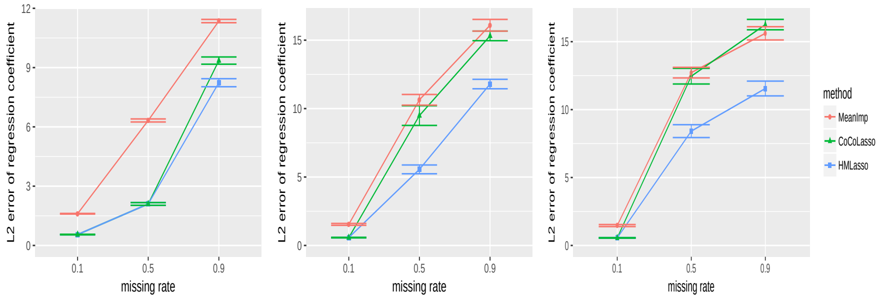

Figure 1: Simulation results with various missing patterns: random pattern (left), column pattern (center), and row column pattern (right).

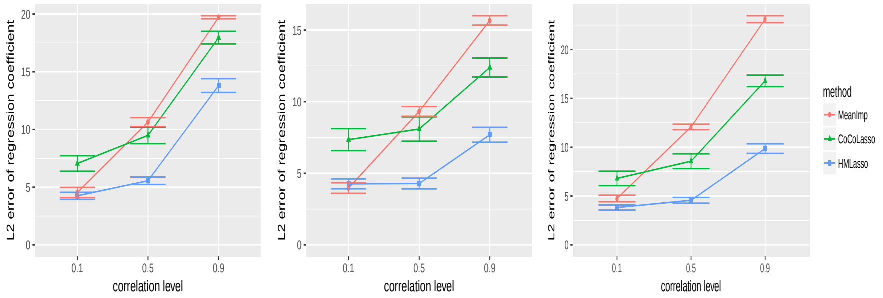

Figure 2: Simulation results with various covariance patterns: uniform (left), autoregression (center), and block (right).

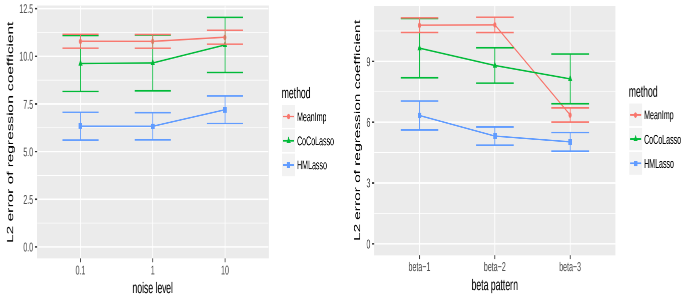

Figure 3: Simulation results with various noise levels (left) and true parameters (right).