Gaussian Graphical Learning via PSD Constraint for Blockwise Missing Multimodal Data

Joseph Graves[1] **Yufeng Liu**[2] **Elio Zhang**[3] **Seong-Tae Kim**[1]

for the Alzheimer's Disease Neuroimaging Initiative*

[1]Mathematics and Statistics Department, North Carolina A & T State University, Greensboro, NC, USA
[2]Department of Statistics, University of Michigan, Ann Arbor, MI, USA
[3]Mathematics Department, Seattle University, Seattle, WA, USA

Abstract

Gaussian graphical models, estimated via precision matrices, are popular for uncovering underlying conditional dependencies among variables, yet the case of multimodal data with blockwise missing remains unaddressed. We propose a novel method, Direct Sparse Graphical Learning for Multimodal Data (DSGL), to handle this issue using only observed data. The DSGL consists of two stages: constructing a pilot estimator for the covariance matrix without imputation as an admissible input for GLASSO; and estimating the DSGL precision matrix via GLASSO with repeated cross-validation tuning. We further propose a thresholded variant to address false-positive edges, a common issue in high-dimensional data. We establish theoretical properties for DSGL with respect to element-wise deviation and its ability to recover the true graphical structure. In simulations and an Alzheimer's Disease Neuroimaging Initiative (ADNI) application, DSGL outperforms imputation-based and other competing approaches, yielding more accurate graph estimation and lower Frobenius-norm error.

1 INTRODUCTION

An emerging paradigm in data analysis is the integration of heterogeneous data modalities to address complex questions. Multimodal data are widely used to improve predictive performance and robustness to noise across applications, from large language models to disease prediction (Zhao et al., 2024; Hurst et al., 2024; Steyaert et al., 2023; Venugopalan et al., 2021). A key challenge with multimodal data is the increased likelihood of blockwise missingness, entire modalities absent for groups of observations, which reduces the stability of imputation-based approaches and raises computational cost (Yuan et al., 2012; Yu et al., 2020). To date, blockwise missing multimodal data research has mainly focused on supervised or semi-supervised learning (Song et al., 2024; Xiang et al., 2014; Yuan et al., 2012; Yu et al., 2020), yet unsupervised problems such as graphical modeling have not been explored.

Gaussian Graphical Models (GGMs) capture conditional dependencies among variables via the non-zero off-diagonal elements of the multivariate Gaussian precision matrix (Fan et al., 2020). Graphical LASSO (GLASSO) and its variants are a popular approach for estimating a GGM precision matrix (Banerjee et al., 2008; Yuan and Lin, 2007; Friedman et al., 2008). Other existing approaches include neighborhood-wise regression (Meinshausen and Bühlmann, 2006; Zhou et al., 2011) and CLIME (Cai et al., 2011, 2016). Furthermore, graphical model estimation has been developed by considering hub detection or joint estimation of multiple GGMs (Tan and Witten, 2015; Sánchez Gómez et al., 2025; Tsai et al., 2024).

Two graphical model issues subsequently arise regarding our study: multimodality and missingness. For multimodal graph estimation, Kolar et al. (2014) proposed a canonical correlation-based estimation method, and Tugnait (2021) introduced a sparse group LASSO. For missingness, most existing research in GGMs is rooted in GLASSO and focuses on providing an alternate estimator to the sample covariance matrix. The standard GLASSO is not directly applicable to blockwise missing multimodal data. Therefore, we need to develop a new estimator for Σ to incorporate GLASSO. Many studies of precision matrix estimation have focused on GLASSO with a single source missing data (Fan et al., 2019; Park et al., 2021; Städler and Bühlmann, 2012). Given the paucity of literature, we are unaware of an ex-

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:1748–1768.

isting study which has developed a GGM with blockwise missing multimodal data.

Another challenge in graphical models is achieving sparsity in the estimator. In the literature, there exists work showing that GLASSO may have a high false positive rate (Liu, 2013; Zhou et al., 2011; Sojoudi, 2016). Therefore, we consider threshold methods, originally introduced as a way to improve sparsity of the covariance matrix estimation (Bickel and Levina, 2008). Thresholding methods in GGM target the sample covariance or other terms that may have some established equivalence with the precision matrix (Li and Gui, 2006; Mazumder and Hastie, 2012). In GLASSO, threshold procedures have been applied to either intermediate steps or the final estimator. Existing papers on thresholding showed that there is a tradeoff between sensitivity, the proportion of true non-zero edges identified, and specificity, the proportion of zero edges identified (Li and Gui, 2006; Laszkiewicz et al., 2021).

In this paper, we propose a novel precision matrix estimation method, Direct Sparse Graphical Learning for Multimodal Data (DSGL) and its threshold variant to address blockwise missingness and reduce false positive edge detection. Our contributions are as follows:

- We construct a pilot positive semi-definite (PSD) guaranteed covariance estimator using only observed data without imputation, motivated by prior research. This pilot estimator serves as critical input to GLASSO and mitigates imputation-induced noise, especially in high-dimensional data. While pilot-estimator ideas have appeared in supervised regression settings (Yu et al., 2020; Yuan et al., 2012), our contribution is an unsupervised, PSD-enforced covariance construction tailored to precision-matrix estimation with blockwise-missing multimodal data.
- We incorporate a thresholding step to our DSGL precision matrix estimator to empirically control the false positive rate of detecting edges by removing spurious small-magnitude entries, improving the graph recovery in the simulation studies.
- We use repeated two-fold cross-validation (CV) to tune both the pilot estimator and the GLASSO penalty. This version of CV uses a stable training and testing split of blockwise missing data for tuning hyperparameters to improve elementwise assessment metrics without relying on the true covariance and precision matrices.
- We provide theoretical work that analyzes the graphical model and covariance matrix under the blockwise missing multimodal scenario. We prove both the theoretical elementwise and graphical convergence of our pilot estimator and the DSGL graphical model estimator to the population level parameters.

1.1 NOTATION

For sample data, let $\boldsymbol{X} \in \mathbb{R}^{n\times p} = (X_1, \ldots, X_p) = (x_1, \ldots, x_n)^T$ denote n observations with p variables generated from a multivariate normal distribution with mean $\vec{0}$ and covariance Σ. Let Σ and $\Omega = \Sigma^{-1}$ denote the true covariance and precision matrices, respectively. The sample mean, a $p \times 1$ column vector, and the sample covariance, a $p \times p$ matrix, without missing are defined as $\bar{\mathbf{x}} = \left(\sum \mathbf{x}_i^T\right)/n$ and $\mathbf{S} = \sum_{i=1}^n \left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^T/(n-1)$.

For vector $b \in \mathbb{R}^p$, we define the L_1 norm, $\|b\|_1 = \sum_{i=1}^{p} b_i$; and the elementwise ∞-norm of b is $\|b\|_\infty = \max(|b_i|)$. For matrix $A \in \mathbb{R}^{n\times m}$, we denote the Frobenius norm as $\|A\|_F = \sqrt{\sum_{i,j}|a_{ij}|^2}$ and the elementwise ∞-norm is defined as $|A|_\infty = \max_{i,j}|a_{ij}|$. Additionally, the L_1 norm is defined as $\|A\|_1 = \max_{1\leqslant j\leqslant n}\sum_{i=1}^{p}|a_{ij}|$; the L_∞ norm is $\|A\|_\infty = \max_{1\leqslant i\leqslant n}\sum_{j=1}^{p}|a_{ij}|$. Finally, we denote eigenvalues of a matrix A as $\nu_1 \geqslant \ldots \geqslant \nu_p$ with the minimum eigenvalue being denoted as $\nu_{min} = \min(\nu_i)$.

Next, we define the modality of variables, the blocks of observations, and blockwise missing. Let M denote the total number of modalities for X, where modality $\mathcal{I}_m$ is an index subset of p_m covariates such that $\mathcal{I}_m = \{m_1, ..., m_{p_m}\}$. Without loss of generality, covariates belonging to the same modality are consecutive, unordered variables. For observation i and modality m, we let $\mathcal{X}_{im} = \{x_{ij} : j \in \mathcal{I}_m \text{ and } x_{ij} \text{ is observed}\}$. We then denote $\mathcal{X}_i = \cup_m \mathcal{X}_{im}$. With this notation, we make the following assumptions:

Assumption 1.1. *For every observation, a modality is either fully observed or entirely missing: $|\mathcal{X}_{im}| \in \{0, p_m\}$. For each pair of modalities $\mathcal{I}_m, \mathcal{I}_t$, there exists i such that $\mathcal{X}_{im} \cap \mathcal{X}_{it} \neq \varnothing$.*

With these assumptions, there are at most 2^M observed missing pattern, defined by which modalities are present or completely missing per observation. Thus, we define $\mathcal{J}(k)$ as the index set of modalities that are the only modalities that are observed, and $\mathcal{B}(k)$ as the subset of observations of X that fit the index set of $\mathcal{J}(k)$.

1.2 GRAPHICAL MODELS AND GLASSO

Covariance matrix entries, $\sigma_{i,j}$, denote the correlation between predictors, but not whether they are dependent or independent ignoring all other variables, called conditional independence. For variables X_i, X_j, they are conditionally independent if $X_i \perp\!\!\!\perp X_j$ given all other predictors. A growing approach to visually representing conditional independence is graphical models. Formally, we can define

a graphical model $\mathcal{G}(V, E)$ as the set of vertices (or nodes) $V = \{1, 2, ..., p\}$ and the edge set $E = \{(i, j) : i < j, \Omega_{ij} \neq 0\}$. In Gaussian graphical model, a set of vertices V represents the variables $\boldsymbol{X}$, and edges represent the conditional dependence between two nodes. So, for variables X_i and X_j, no edge exists between their nodes if and only if X_i and X_j are conditionally independent given the other variables (Hastie et al., 2009). The precision matrix Ω of multivariate normal random variables is a direct representation of the conditional dependence between variables, i.e., variables X_i, X_j such as:

$$(i, j) \notin E \iff X_i \perp\!\!\!\perp X_j | X_{-(i,j)} \iff \Omega_{ij} = 0.$$

The Graphical LASSO was proposed to estimate Ω, motivated by converting the log-likelihood optimization problem into a lasso-based approach (Meinshausen and Bühlmann, 2006; Banerjee et al., 2008). The goal of GLASSO is to maximize the following log likelihood function via estimator $\hat{\Sigma}^{-1}$ s.t.

$$\hat{\Sigma}^{-1} = \arg\max_{\Theta > 0} \log\det(\Theta) - tr(S\Theta) - \lambda\|\Theta\|_1. \quad (1)$$

As seen in supervised learning problems, providing a sufficient substitute for S is an important step. Earlier work in this field focused on using as much of the observed data as possible (Städler and Bühlmann, 2012; Kolar and Xing, 2012). This strategy further developed by imposing additional weights and stipulations based on some rule of missingness (Loh and Tan, 2018; Park et al., 2021). Common issues include ensuring that the replacement estimator for S is both unbiased and positive (semi)definite, which is a challenge in the presence of blockwise missing.

1.3 PRELIMINARY COVARIANCE CONSTRUCTION WITH BLOCKWISE MISSING MULTIMODAL DATA

Figure 1: Blockwise Missing Multimodal Data

Let $\tilde{\mathbf{X}}$ be the multimodal data matrix with blockwise missingness under Assumption 1.1, which is illustrated in Figure 1. The standard covariance estimation is infeasible for $\tilde{\mathbf{X}}$. Hence, we use a preliminary covariance estimator, $\tilde{\Sigma}$, proposed by Yu et al. (2020), as follows:

1. For predictors j and r, let $\mathcal{I}_j$ and $\mathcal{I}_r$ be their respective modalities.
2. Define $\mathcal{B}_{jr} = \cup_k \{\mathcal{B}(k) : \mathcal{I}_j, \mathcal{I}_r \in \mathcal{J}(k)\}$, and $n_{jr} = |\mathcal{B}_{jr}|$.
3. Then $\tilde{\Sigma}$ is defined as:

$$\tilde{\Sigma} = (\tilde{\sigma}_{jr})_{j,r=1,...,p}, \quad (2)$$

where

$$\tilde{\sigma}_{jr} = \sum_{i \in \mathcal{B}_{jr}} x_{ij} x_{ir} / (n_{jr}). \quad (3)$$

The key strength of $\tilde{\Sigma}$ is that its construction is based on all existing data and imposes a weight on covariance elements based on the number of existing elements for a pair of predictors without removing or imputing missing elements. For example, in Figure 1, $\mathcal{B}_{11}$ uses data in all blocks because Modality 1 does not have any missing, $\mathcal{B}_{12} = \mathcal{B}_{21}$ uses data in Blocks 1 and 2, $\mathcal{B}_{13} = \mathcal{B}_{31}$ uses Blocks 1 and 3, and $\mathcal{B}_{23} = \mathcal{B}_{32}$ uses only Block 1 data. The initial estimator $\tilde{\Sigma}$ is unbiased, but it is still not guaranteed to be positive semidefinite. To resolve this issue, we represent $\tilde{\Sigma}$ as an $M \times M$ symmetric block matrix, such that each element $\tilde{\Sigma}_{jt}$ is a $p_j \times p_t$ for $j, t = 1, \cdots, M$. Using this scheme, define the intramodality block diagonal matrix

$$\tilde{\Sigma}_D = DIAG(\tilde{\Sigma}_{11}, \ldots, \tilde{\Sigma}_{MM}), \quad (4)$$

and the intermodality block matrix

$$\tilde{\Sigma}_C = \tilde{\Sigma} - \tilde{\Sigma}_D. \quad (5)$$

These two preliminary estimated matrices will be utilized to ensure the positive semi-definiteness of a pilot covariance estimator.

2 METHODOLOGY

2.1 PILOT COVARIANCE ESTIMATOR CONSTRUCTION

As we discussed in Section 1.3, the preliminary covariance estimate $\tilde{\Sigma}$ does not guarantee a positive semi-definiteness, which is not an admissible input for the GLASSO optimization. Therefore, we define the positive definite pilot covariance estimator, $\hat{\Sigma}_\alpha$:

$$\hat{\Sigma}_\alpha := \tilde{\Sigma}_D + \alpha\tilde{\Sigma}_C + \nu \cdot I_p, \quad (6)$$

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at: `adni.loni.usc.edu`

where $\nu = \left(|\nu_{min}(\tilde{\Sigma}_D + \alpha\tilde{\Sigma}_C)| + \delta\right) \cdot I(\nu_{min} < 0)$, which is asymptotically ignorable for small $\delta > 0$. We then normalize this quantity as in 2.1.

Definition 2.1 (Normalized Positive Definite Covariance Estimator)**.**

$$\hat{\Sigma}_{\nu,\alpha} := \frac{1}{1+\nu}\tilde{\Sigma}_D + \frac{\nu}{1+\nu}I_p + \frac{\alpha}{1+\nu}\tilde{\Sigma}_C. \quad (7)$$

The normalization process is useful to simply the proof that $\hat{\Sigma}_{\nu,\alpha} \to \Sigma$ under sufficient conditions, as seen in Theorem 3.1.

Although our normalized pilot covariance estimator $\hat{\Sigma}_{\nu,\alpha}$ has the same basis as preceding research, there are notable differences. First, the tuning parameter $\alpha \in (0, 1]$ controls the weight between the intramodality and intermodality portions of the preliminary estimator. Theoretically, we require α to satisfy the following condition: $\alpha = O\left(\sqrt{\log(p)/\min_{j,r} n_{jr}}\right)$. This type of pilot estimator is a different version of preceding covariance estimators using all data without imputation (Kolar and Xing, 2012; Yu et al., 2020). Therefore, we capture the resulting imbalance of observable data between different modalities.

Yu et al. (2020) uses a linear combination of $\tilde{\Sigma}_C$ and $\tilde{\Sigma}_D$ where any resulting non-positive definite combination of the two matrices is discarded in their algorithm. Also, there is an additional weight allowing for $\tilde{\Sigma}_C$ to have more impact in the resulting regression problem than $\tilde{\Sigma}_D$. By reducing the dimensionality of our estimator and adding the minimum eigenvalue, we can test every possible combination of α and the GLASSO tuning parameter, λ, in reasonable computing time.

2.2 DSGL

Algorithm 1 outlines how to obtain the precision-matrix estimate from the sample blockwise multimodal data with missingness, $\tilde{\boldsymbol{X}}$, using the pilot covariance estimates in Eqs. 6 and 7. This algorithm produces GLASSO precision-matrix estimates for specified tuning parameters α and λ, with the optimal values determined via a tuning procedure.

In previous research, tuning methods for λ focused on either Bayesian Information Criterion (BIC) or Cross-Validation. The BIC penalty adds degrees of freedom based on the number of detected edges to enforce sparsity (Yuan and Lin, 2007). Graphical model estimation in the MCAR that used BIC yielded far higher sparsity in their estimations (Städler and Bühlmann, 2012; Kolar and Xing, 2012). A K-fold cross-validation (CV) approach typically trains the graphical model estimation on $K-1$ folds and evaluates the estimate against the remaining fold. In terms of the missing data setting, cross-validation typically yielded better estimates in terms of a specified loss function but did not have as much sparsity as the BIC penalty (Städler and Bühlmann, 2012). Preliminary runs of DSGL highlighted a tendency to yield the smallest λ candidate value possible. BIC penalty inversely led to few correct edges detected. CV approaches therefore become the more attractive option to use as much data as possible, but the standard K-fold approach raises concerns about stability in the blockwise missing multimodal setting.

Algorithm 1 DSGL-Main

Input: Data matrix $\tilde{\boldsymbol{X}} \in \mathbb{R}^{n\times(P_1,\ldots,P_M)}$.
Construct $\tilde{\Sigma}$, $\tilde{\Sigma}_D$ and $\tilde{\Sigma}_C$ as in Eqs. 2 - 5.
given $0 \leqslant \alpha < 1$, $0 < \lambda < 1$, $\delta > 0$
Construction of Eq. 7
$\hat{\Sigma}_\alpha \leftarrow \tilde{\Sigma}_D + \alpha\tilde{\Sigma}_C$
$\nu \leftarrow$ minimum eigenvalue of $\hat{\Sigma}_\alpha$
if $\nu \leqslant 0$ **then**
  $\nu \leftarrow |\nu| + \delta$
  $\hat{\Sigma}_{\nu,\alpha} \leftarrow \frac{1}{1+\nu}\tilde{\Sigma}_D + \frac{\nu}{1+\nu}I_p + \frac{\alpha}{1+\nu}\tilde{\Sigma}_C$
else if $\nu > 0$ **then**
  $\hat{\Sigma}_{\nu,\alpha} \leftarrow \hat{\Sigma}_\alpha$
end if
Run GLASSO on $\hat{\Sigma}_{\nu,\alpha}$
$\hat{\Omega}_{\lambda,\alpha} \leftarrow GLASSO(\hat{\Sigma}_{\nu,\alpha}, \lambda)$
return: $\hat{\Omega}_{\lambda,\alpha}$

Therefore, to find the optimal tuning parameters α and λ, we apply a K repeated, two-fold CV. The folds are constructed by randomly splitting $\boldsymbol{X}$ so that each fold has an even number of observations for each observed missing modality combination β_r. We construct the preliminary estimators for each split, and then fix one as the reference estimator for Σ we call $\tilde{\Sigma}^1$. For the other split, we apply Algorithm 1 for each candidate (α, λ) pair. We then define the following loss-function, an empirical version of the log-likelihood loss function used during GLASSO:

$$l_1(\hat{\Omega}, \tilde{\Sigma}) = -\log\det(\hat{\Omega}) + \text{tr}(\tilde{\Sigma}\hat{\Omega}). \quad (8)$$

Although, Friedman et al. (2008) used 10-fold CV using a log-likelihood function with the $\lambda\|\Omega\|_1$, that used complete observations. The Städler and Bühlmann (2012) version of the CV-score uses a version of the log-likelihood function based on missing data without the L_1 penalty. Therefore, Eq. 8 is based on the estimate $\hat{\Omega}$ of one fold evaluated against $\tilde{\Sigma}^1$. We repeat the process treating the other split as the reference estimator, and apply Algorithm 1. Let the corresponding k-fold CV value for (λ, α) be the sum of the two l_1 values s.t. $CV^k(\lambda, \alpha) = l_1(\hat{\Omega}^1, \tilde{\Sigma}^2) + l_1(\hat{\Omega}^2, \tilde{\Sigma}^1)$. Therefore, our optimal parameter is chosen to satisfy the following:

$$(\lambda^*, \alpha^*) = \arg\min \sum_{k=1}^{K} CV^k(\lambda, \alpha).$$

In the event of a tie, we prioritize the largest α value which promotes higher intermodal influence then the smallest λ to ensure more edges are detected, as we expected to improve sparsity with thresholding. See Algorithm 2 for additional details. As seen in our simulation runs, the repeated CV was able to avoid border solutions for our optimal λ while satisfying the asymptotic properties of α.

Algorithm 2 DSGL - Tuning

Input: Data matrix $\tilde{\boldsymbol{X}} \in \mathbb{R}^{n \times (P_1,\ldots,P_M)}$.
given $\alpha = (\alpha_1, \ldots \alpha_A)$, and $\lambda = (\lambda_1, \ldots \lambda_L)$
$CV \in \mathbb{R}^{A \times L}$ initialized as zero matrix.
for Iteration $1, \ldots, k, \ldots, K$ **do**
  $CV^k \in \mathbb{R}^{A \times L}$ as a zero matrix .
  Set random sampling seed based on fold k
  for Each $\mathcal{B}_{jr} \in \mathcal{B}$ **do**
    Define $\mathcal{B}^1_{jr}, \mathcal{B}^2_{jr}$ as a even, randomly sampled splits of $\mathcal{B}_r$
  end for
  $\tilde{\boldsymbol{X}}^1 \leftarrow (\mathcal{B}^1_1, \ldots, \mathcal{B}^1_R)$, $\tilde{\boldsymbol{X}}^2$ defined similarly.
  $\tilde{\Sigma}^1 = \tilde{\Sigma}(\tilde{\boldsymbol{X}}^1)$, similarly for $\tilde{\Sigma}^2$
  for α_a, λ_l **do**
    $\hat{\Omega}^1 \leftarrow DSGL(\boldsymbol{X}^1)$, $\hat{\Omega}^2 \leftarrow DSGL(\boldsymbol{X}^2)$
    Compute $L_1 = l_1(\hat{\Omega}^1, \tilde{\Sigma}^2)$, $L_2 = l_1(\hat{\Omega}^2, \tilde{\Sigma}^1)$.
    $(CV^K)_{a,l} = L_1 + L_2$
  end for
end for
$CV = \sum CV^K$, let $CV_{MIN} = \min CV$
Choose (λ^*, α^*) corresponding to CV_{MIN}
return: (λ^*, α^*) and $\hat{\Omega}(\lambda^*, \alpha^*)$

2.3 THRESHOLDING

To mitigate excessive false-positive edges in GGM estimates for high-dimensional data, we adopt a hard thresholding procedure based on the chosen (λ^*, α^*) from the training process. We define the following, let $C \in (0, 1]$, and $\tau = C\lambda^*$, and $\hat{\Omega}$ be the corresponding DSGL precision estimator associated with those parameters. Then, our threshold version of the DSGL estimator is proposed as:

$$[\hat{\Omega}^\tau]_{i \neq j} = \hat{\Omega}_{ij} I\left(|\hat{\Omega}_{ij}| \geqslant \tau\right) \tag{9}$$

where the threshold is a hard threshold as in Bickel and Levina (2008). Other thresholds such as soft threshold or adaptive threshold are also possible (Cai and Yuan, 2012; Donoho et al., 1995). The threshold variant of the DSGL precision matrix is estimated using Algorithm 3. More recently, Laszkiewicz et al. (2021) applied thresholding to their relatively novel adaptive validation procedure for tuning λ after providing a graphical model estimation. They proceed to make a recommendation based on the optimal tradeoff between the sensitivity and the positive predictive value defined in our simulation report. Therefore, in our simulation settings, we establish our own decision rule to recommend an optimal threshold parameter.

One potential drawback is that this hard-thresholded version may not be positive definite, despite the initial estimate being guaranteed to be so from GLASSO. In that scenario, we modify $\hat{\Omega}^\tau$ based on the minimum eigenvalue to ensure positive definiteness.

Algorithm 3 DSGL: Thresholding

Input: (λ^*, α^*) and $\hat{\Omega}(\lambda^*, \alpha^*)$ from Algorithm 2.
Define $\vec{C} = [C_1, \ldots, C_{m-1}, 1]$ as a range of values modifying λ^*
for $C_1, \ldots, C_m$ **do**
  $\tau = C_i\lambda^*$
  Construct $\hat{\Omega}^\tau$ as outlined in Eq. 9.
  if $\nu_{min}(\hat{\Omega}^\tau) < 0$ **then**
    $\hat{\Omega}^\tau \leftarrow \hat{\Omega}^\tau + |\nu_{min}| I_p$
  end if
  Construct confusion matrix based on upper triangular matrices of $\hat{\Omega}^\tau$ with Ω as the reference.
  Choose τ^* based on the decision rule in Eq. 10
end for
return: τ^* and $\hat{\Omega}(\lambda^*, \alpha^*, \tau^*)$.

3 THEORETICAL FINDINGS

In this section, we establish three main theorems concerning the precision-matrix estimator in the setting of blockwise missing multimodal data: tail condition, elementwise convergence, and graphical recovery. For notational convenience, let $\hat{\Sigma} := \hat{\Sigma}(\lambda^*, \alpha^*)$ and $\hat{\Omega} := \hat{\Omega}(\lambda^*, \alpha^*)$.

3.1 TAIL CONDITIONS OF $\hat{\Sigma}$

Now, we show elementwise convergence of $\hat{\Sigma}_{\nu,\alpha}$ to Σ and that of the resulting DSGL estimate $\hat{\Omega}$. We adapt the tail conditions from Ravikumar et al. (2011) to describe the estimator $\hat{\Sigma}$. These tail conditions were further outlined by Kolar and Xing (2012) and Yu et al. (2020) in the context of missing data.

We state the following assumptions necessary to prove our estimator $\hat{\Sigma}_{\mu,\alpha}$ satisfy our chosen tail condition.

Assumption 3.1 (Sub-Gaussian Distribution)**.** *Random variable X_i is Sub-Gaussian if there exists a constant L s.t.*

$$E[\exp[tX_j]] \leqslant \exp\left[L^2t^2/2\right].$$

Assumption 3.2 (Sample size Assumption)**.** $tn \geqslant 6\log(p^\varphi)$*, where* $0 < t < 1$ *denotes the smallest proportion of observations observed for each pair of modalities, and* $\varphi > 2$.

Assumptions 3.1 and 3.2 are based on the assumptions made in DISCOM. Assumption 3.2 is an extended version of the condition of Theorem 1 of DISCOM to account for a more generic tail condition function, where asymptotic behavior depends on the entire sample size, n, instead of the smallest block size.

We define $Z_i = X_i/\sqrt{1+\nu}$ to be the transformation of X_i. Immediately we have that $Z_i \sim \mathrm{N}(0, \Sigma_Z)$ such that $\Sigma_Z = \frac{1}{1+\nu}\Sigma$. Immediately, we have that $Z_i/(\Sigma_Z)_{ii}$ follows a sub-Gaussian distribution, a necessary condition for our theoretical findings. We then define $\tilde{\Sigma}_Z = \frac{1}{1+\nu}\tilde{\Sigma}$.

Theorem 3.1 (Tail Condition of DSGL)**.** *Let $\varphi > 2$ and $tn > 6\log(p^{\varphi})$, and $Z_i/(\Sigma_Z)_{ii}$ be sub-Gaussian for parameter L. Define $\hat{\Sigma}_{\nu,\alpha}$ as in (7). If $\delta^* = A\sqrt{\log(p^{\varphi})/(tn)}$ where $A = 8(1+4L^2)\sqrt{6(1+\nu)^{-1}}$. Then:*

$$P\left(|(\hat{\Sigma}_{\nu,\alpha} - \Sigma_Z)_{ij}| \geqslant \max\left\{(1+\nu)^{-1}, \delta^*\right\}\right) \leqslant 4p^{-3\varphi}.$$

Remark: Theorem 3.1 states that Z_i satisfies the tail condition for $f(n,\delta) = \frac{1}{4}\exp(c_* n\delta^2)$, with $c_* = (1+\nu)[128(1+4L^2)^2]^{-1}$ and $\delta \in [0, \frac{8}{\sqrt{1+\nu}}(1+4L^2)]$ with δ^* satisfying the inequality. The proof for this theorem is an intermediate step in proving Theorem 2 of Yu et al. (2020), outlined further in the supplement.

Corollary 3.1.1. *Let $\varphi^* = 3\varphi/t - \log_p(4)$. Then $f(n,\delta^*) = \exp\left[-3\log(p^{\varphi})/t\right]/4 = p^{\tau^*}$ such that*

$$\bar{\delta}_f(n, p^{\varphi^*}) = \sqrt{\log(4p^{\varphi^*})/(c_* n)} = \delta^*.$$

Corollary 3.1.1 establishes that δ^* can be expressed as the maximum δ as the inverse of a tail function. The proof is established by straightforward algebraic manipulation. The main purpose of this is to establish a lower bound on the sample size of n.

3.2 ELEMENTWISE CONVERGENCE OF $\hat{\Omega}$

We establish the following notation from Ravikumar et al. (2011). Let $\kappa_\Sigma = \|\Sigma\|_\infty$ be the infinity norm of the true covariance matrix. Then define $\Gamma^* = \Omega^{-1} \otimes \Omega^{-1}$ where an entry of Γ^* is denoted as $\Gamma^*_{(j,k),(l,m)}$. $S = S(\Omega) = \{E(\Omega) \cup (1,1) \cup \ldots \cup (p,p)\}$, S^C is the complement of S. Then s is the order of the edge set with $|S| = |E(\Omega)| = s + p$, and d is the maximum degree of Ω. Then $\Gamma^*_{SS} \in \mathbb{R}^{(s+p)\times(s+p)}$ is Γ^* indexed by its edge sets, and its corresponding norm is $\kappa_{\Gamma^*} = \|(\Gamma^*_{SS})^{-1}\|_\infty$. Additionally, we denote $\kappa_{max} := \max\{\kappa_\Sigma\kappa_\Gamma, \kappa_\Sigma^3\kappa_\Gamma^2\}$ and $\bar{\gamma} = (1 + 8/\gamma)$.

We state a new assumption regarding the effect of zeros on the edge-set:

Assumption 3.3 (Irrepresentability Condition (Ravikumar et al., 2011))**.** *There exists $\gamma \in (0,1]$ s.t.*

$$\max_{e \in S^c} \|\Gamma_{eS}(\Gamma_{SS})^{-1}\|_1 \leqslant 1 - \gamma.$$

Theorem 3.2 (DSGL Elementwise Convergence)**.** *Let a random distribution satisfy Assumption 3.3 for some parameter γ and satisfy the tail condition $\mathcal{T}(f, v_*)$. $\hat{\Omega}$ is the unique solution to DSGL with $\lambda^* = (8/\gamma)\delta^*$ where $\varphi > 2$ and $\delta^* = A'\sqrt{\log(p^{\varphi^*})/(tn)}$, where $\tau^* = 3\varphi/t - \log_p 4, \varphi > 2$, with $A' > A$, A is defined as it was in Theorem 3.1. Also let d denote the maximum degree of Ω. Let the sample size be lower bounded such that:*

$$n > 3t^{-1}Cd^2\bar{\gamma}^2\log(p^{\varphi}),$$

where $C = (1+\nu)^{-1}[48\sqrt{2}(1+4L^2)\kappa_{max}]^2$ and

$$48\kappa_\Sigma^3\kappa_\Gamma^2(1+\gamma/8)^2 d\lambda^* \leqslant \gamma.$$

Then $\hat{\Omega}$ satisfies the following conditions with probability greater than $1 - 1/p^{(\varphi^*-2)} \to 1$.

1. *$\hat{\Omega}$ is elementwise l_∞ bounded: $\|\hat{\Omega} - \Omega\|_\infty \leqslant \{2\bar{\gamma}\kappa_{\Gamma^*}\}\bar{\delta}_f(n, p^{\varphi^*})$.*
2. *$E(\hat{\Omega}) \subset E(\Omega)$ and furthermore contains all edges of $E(\Omega)$ where $|\Omega_{ij}| > \{2\bar{\gamma}\kappa_{\Gamma^*}\}\bar{\delta}_f(n, p^{\varphi^*})$.*

Remark: In terms of λ^*, we can express the bound on the maximum deviation of $\hat{\Omega}$ as $C_{\lambda^*} = \frac{1}{4}\bar{\gamma}\kappa_\Gamma\lambda^*$.

3.3 GRAPHICAL RECOVERY

To determine how well our estimator $\hat{\Omega}$ recovers the true edge set $E(\Omega)$ of Ω, we use the following terms:

$$\omega_{\min} := \min_{(i,j)\in E(\Omega)} |\Omega_{ij}|,$$

and the event

$$\mathcal{M}(\hat{\Omega}, \Omega) := \left\{\mathrm{sign}\{\hat{\Omega}_{ij}\} = \mathrm{sign}\{\Omega_{ij}\} \forall\ i,j\ \in E(\Omega)\right\}.$$

We construct the following theorem based on Theorem 2 of Ravikumar et al. (2011):

Theorem 3.3 (DSGL Graphical Recovery)**.** *Let a random distribution satisfy Assumption 3.3 for some parameter γ and satisfy the tail condition $\mathcal{T}(f, v_*)$. $\hat{\Omega}$ is the unique solution to Graphical LASSO with $\lambda^* = (8/\gamma)\delta^*$ where $\varphi > 2$ and δ^* the same as in Theorem 3.2. Let the sample size be lower bounded such that:*

$$n > \bar{n}_f\left(1/\max\{2\kappa_\Gamma(\bar{\gamma}\omega_{min})^{-1},\ v_*, 6\bar{\gamma}d\kappa_{max}\}, p^{\varphi^*}\right).$$

Then our estimated graph selection is consistent as $p \to \infty$:

$$\mathcal{P}(\mathcal{M}(\hat{\Omega}, \Omega)) \geqslant 1 - p^{\varphi^*-2} \to 1.$$

In terms of graphical recovery and thresholding, both Bickel and Levina (2008) and Laszkiewicz et al. (2021) provide conditions regarding the threshold version of a precision matrix estimation captures the edge set. The latter is more general, as it does not specify a particular sparsity requirement that the population level covariance matrix satisfies.

4 EXPERIMENTAL RESULTS

4.1 SIMULATIONS

We generate the true precision matrix based on the Erdös-Renyi graph, with edges drawn elementwise as outlined below:

$$P(\boldsymbol{B}_{ij} = .5) = \begin{cases} .075 & p_i, p_j \in M_k \\ .05 & p_i \in M_m, p_j \in M_k, m \neq k \end{cases}$$

Add δI_p s.t. $\Omega = \boldsymbol{B} + \delta I_p$ to ensure positive-definite Ω. Choose $\delta = c \min \text{EV}(\boldsymbol{B})$ where $c > 1$ is chosen to satisfy Assumption 3.1 (Städler and Bühlmann, 2012; Park et al., 2021). We compute the inverse of the true precision matrix, $\Sigma = \hat{\Omega}^{-1}$. Next, we generate $\boldsymbol{X} = (X^{(1)}, ..., X^{(p)}) \in \mathbb{R}^{n \times p}$ with a multivariate normal distribution with mean $\vec{0}$ and covariance Σ.

We then define blockwise missingness $\tilde{\boldsymbol{X}}$ by introducing a predetermined missing pattern. These missing data patterns are determined by the percentage of observed data per modality (m_1, m_2, m_3), the percentage of data per observational block (B_1, B_2, B_3, B_4), and the minimum number of shared observations between modalities $\min n_{M_i, M_j}$. We test a fixed sample size of $n = 400$ and predictor size $p_{300} = (100, 100, 100)$. We therefore test the following missing pattern, based on the pattern outlined in Figure 1:

- Missing Pattern 1: $(100, 50, 50)$; Block Distribution: even; and $\min n_{M_i, M_j} = .25N$.
- Missing Pattern 2: $(100, 90, 90)$. Block Distribution: $(85, 5, 5, 5)$. $\min n_{M_i, M_j} = .85N$

We compute the $\tilde{\Sigma}$ estimate of the training data with the missing data imposed,and then collect the largest non-diagonal element in magnitude across all thirty repetitions, which we label λ_{max}. We then take a sequence of values up to λ_{max}. For α we test $\hat{\alpha} \in [.05, 1]$ in increments of .05 using Algorithm 2. The GLASSO step is done using the R package **glasso** (Friedman et al., 2019).

Now, using 30 iterations of simulation, we compare DSGL, obtained via the data generating steps above and Algorithms 1-3, to the Soft Imputation, complete-case only, *i.e.*, computing a precision matrix estimate from only the complete observations, and more advanced imputation methods like MICE (Hastie et al., 2015; Mazumder et al., 2010; van Buuren and Groothuis-Oudshoorn, 2011). In all these scenarios, unless otherwise stated, we fix $\alpha = 1$ as missing values have been imputed.

We evaluate DSGL with two types of metrics: (1) model fit via selected tuning parameters and Frobenius-norm error between the estimated and true precision matrices; and (2) edge detection via a confusion matrix computed from the upper-triangular entries of the estimated precision matrix as follows:

As highlighted in Algorithm 3, for thresholding, we create a confusion matrix based on the correctly identified nonzeroes and zeroes of the upper triangular matrix compared to the true Ω. Then, we record our graphical assessment metrics as defined earlier. Without loss of generality, we define $OPT(\tau) = (TPR + PPV)(\tau)$ as the sum of the TPR and PPV for an estimate $\hat{\Omega}$ thresholded at a specific value of τ. In general, per the nature of thresholding, as $C \to 1$, the sensitivity is monotonically decreasing, and the specificity is monotonically increasing. It follows that we want to improve the PPV of our estimate without sacrificing the overall TPR, we define $TPR \geqslant \hat{\tau}_0$ as the tolerance value for TPR. Referring to Algorithm 3, we define the following decision rule for $\hat{\tau}^*$:

$$\tau^* := \underset{TPR \geqslant \hat{\tau}_0}{\arg\max} OPT(\tau). \tag{10}$$

In the event $\overline{TPR} < \hat{\tau}_0$, we choose not to threshold.

Tables 1, 2 show the simulation results comparing DSGL with competing methods for $n = 400, p = 300$ with 30 iterations, where the cell values indicate the mean and standard deviation of the assessment metrics. DSGL satisfied the asymptotic conditions of α necessary for our theoretical work, with $\alpha \to 1$ with more data available. Furthermore, particularly in the first missing pattern scenario, DSGL yielded the highest TPR while maintaining high Specificity in both the thresholded and non-thresholded versions. With hard thresholding, there was the possibility of deleting correctly identified edges, particularly in the intermodal portions of $\hat{\Omega}$. However, thresholding in both missing pattern scenarios improved the F_1 and PPV significantly while preserving a high sensitivity. MICE in particular suffered poorly in terms of edge recovery. It also had the longest runtime by a magnitude of days compared to the other methods.

With Missing Pattern 2, DGSL's results remain the same with an improvement in standard deviation. The comparison methods improve, as a significant amount of data is observable. MICE and Soft Imputation in particular yielded a higher PPV than the non-thresholded DSGL. Despite this, DSGL maintained an advantage in terms of TPR and outperformed the competing methods in PPV once thresholding was introduced, meaning that even if thresholding the other methods, DSGL will still maintain a clear advantage.

Our next set of simulations focus on the Gaussian Nonparanormal (NPN) distribution, which weakens the sub-Gaussian assumption by introducing a nonparametric extension via a differentiable transformation (Liu et al., 2009, 2012). Notably, this distribution still preserves the equivalence between zeros in the precision matrix and the conditional independence between two covariates, allowing for the use of precision matrix estimation methods to estimate the graphical model (Liu et al., 2009, 2012; Wang et al., 2014). We extend this notion, by introducing heterogeneity, transforming each modality by a different func-

Table 1: The mean(sd) of model fit and edge detection metrics for Random Graph, $n = 400, p = 300$, Missing Pattern (MP) 1. TPR = True Positive Rate; TNR = True Negative Rate; PPV = Positive Predicted Value.

Method	$\hat{\lambda}^*, \hat{\alpha}^*/\hat{\tau}^*$	$\|\hat{\Omega} - \Omega\|_F$	TPR	TNR	F_1	PPV
DSGL	[0.13(0.01), .65(0.01)]	**3.99(0.08)**	**85.94(1.89)**	97.96(0.48)	50.60(3.57)	35.99(3.58)
DSGL, thresh	$\tau = .1\lambda^*$	4.02(0.08)	72.31(2.73)	99.44(0.18)	**67.43(2.45)**	**63.52(4.77)**
MICE	[0.68(0.15), 1]	5.08(0.08)	8.08(4.46)	**99.90(0.07)**	13.35(6.94)	58.59(14.67)
Comp Only	[0.37 (0.01), 1]	5.37(0.04)	48.26(1.91)	99.63(0.03)	54.60(1.49)	62.93(1.68)
Soft Imput	[0.23(0.01), 1]	4.66(0.08)	72.56(2.50)	99.26(0.07)	63.39(0.93)	56.36(1.81)

Table 2: The mean(sd) of model fit and edge detection metrics for Random Graph, $n = 400, p = 300$, MP 2.

Method	$\hat{\lambda}^*, \hat{\alpha}^*/\hat{\tau}^*$	$\|\hat{\Omega} - \Omega\|_F$	TPR	TNR	F_1	PPV
DSGL	[0.15(0), 1(0)]	**3.98(0.03)**	**87.57(1.16)**	97.93(0.08)	50.71(0.94)	35.69(0.90)
DSGL, thresh	$\tau = .1\lambda^*$	4.00 (0.04)	76.54(1.13)	**99.53(0.03)**	**71.99(.96)**	**67.96(1.24)**
MICE	[0.28(.01), 1]	4.98(0.07)	62.03(2.62)	99.49(0.04)	61.34(1.53)	60.73(1.52)
Comp Only	[0.18(0.01), 1]	4.29(0.11)	81.47(2.42)	98.78(0.16)	59.47(1.76)	46.94(2.87)
Soft Imput	[0.20(0.01), 1]	4.43(0.06)	78.27(1.69)	99.02(0.09)	61.81(1.32)	51.12(1.91)

tion and introducing blockwise missingness. We use the same Ω from the multivariate normal scenario. We generate $\boldsymbol{Z}_1, \boldsymbol{Z}_2 \sim N(0, \Sigma) \in \mathbb{R}^{n \times p}$, then transform modality i by $f_i(Z)$, where $f_1(Z) = Z; f_2(\boldsymbol{Z}_1, \boldsymbol{Z}_2) = \boldsymbol{Z}_1 + \boldsymbol{Z}_2$, and $f_3(\boldsymbol{Z}) = \text{sign}\{\boldsymbol{Z}\} * |\boldsymbol{Z}|^{(1/3)}$. The final data is $\boldsymbol{X}_{NPN} \sim NPN(0, \Sigma, f)$, the Gaussian copula estimation using a shrunken ECDF transformation. This step involved the **huge.npn** function from the **huge** package (Zhao et al., 2012; Jiang et al., 2026).

Tables 3, 4 show the simulation results of DSGL with competing methods for $n = 400, p = 300$ with the same iterations as in the previous simulation. All methods uniformly perform worse than the standard Gaussian multivariate simulations in F_1 and Positive Predictive Value. DSGL maintains the highest true positive rate across all methods, while MICE suffers the lowest edge recovery in Missing Pattern 1. Soft Imputation had similar performance metrics to DSGL, but outperformed in F_1 and Positive Predictive Value, even after thresholding DSGL. In Missing Pattern 2, all methods improved as expected. Furthermore, we note that $\hat{\alpha}$ approaches 1 as more data is introduced. DSGL still had the highest True Positive Rate, and outperformed all other methods in the other categories in the thresholded version. We are interested in achieving stronger nonparametric convergence in the future.

4.2 REAL DATA APPLICATION

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). There are 93 MRI, 93 PET, and 5 Cerebrospinalfluid (CSF) input variables (Mueller et al., 2005). Because of its blockwise missing multimodal structure, this dataset is popular for real data application in corresponding analysis (Yu et al., 2020; Xue and Qu, 2021). We recreate the same repeated cross validation approach by randomly creating training and test data splits, in which we find the optimal $\hat{\lambda}^*$ and $\hat{\alpha}^*$ parameters. The true graphical structure of the ADNI data is unknown, so we take a stability approach motivated by Basu et al. (2018); Liu et al. (2010). Let $\boldsymbol{X}^1_{test}, \ldots, \boldsymbol{X}^{30}_{test}$ denote the test data splits of ADNI and $E(\hat{\Omega}^1), \ldots, E(\hat{\Omega}^{30})$ denote the respective edge set approximates from DSGL. We therefore define the stability for a potential edge e_{ij}, ie. the graphical connection between covariates i and j as:

$$sta(e_{ij}) = \frac{1}{30} \sum_{k=1}^{30} \mathbb{I}\left(e_{ij} \in E(\hat{\Omega}^k)\right).$$

Let $\hat{s}_0$ be a predetermined stability value and $E(\hat{\Omega})$ as the composite edge set across the thirty iterations. The procedure is as follows, for candidate edge e_{ij}, if $sta(e_{ij}) \geqslant \hat{s}_0$, we report $e_{ij} \in E(\hat{\Omega})$. Otherwise, $e_{ij} \in E(\hat{\Omega})^C$. As an example, as seen in Figure 2a, for $\hat{s}_0 = 1$, a $e_{ij} \in E(\hat{\Omega})$ only if $e_{ij} \in E(\hat{\Omega}^k)$ for all k.

Figure 2 shows sample graphs of the edges consistently detected by all data splits. We note that in Figure 2a, $\hat{s}_0 = 1$ with no thresholding, few intermodal edges between the MRI and PET modalities are identified, and CSF is isolated. The visual showed fairly weak intermodal connections if

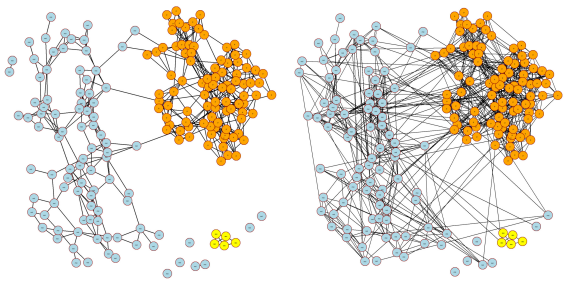
Table 3: The mean(sd) of model fit and edge detection metrics for NonParanormal Data: $n = 400, p = 300$, MP 1

Method	$\hat{\lambda}^*, \hat{\alpha}^*/\hat{\tau}^*$	$\|\hat{\Omega} - \Omega\|_F$	TPR	TNR	F_1	PPV
DSGL	[0.12(0.00),0.65(0.00)]	3.93(0.03)	**77.53**(1.81)	98.77(0.05)	39.75(1.06)	26.73(0.85)
Thresholded	$\tau = .1\lambda^*$	3.93(0.03)	72.56(1.84)	99.10(0.03)	44.17(.98)	31.76(0.80)
Comp Only	[0.36(0.01), 1]	4.14(0.05)	49.42(3.01)	99.49(0.05)	41.59(1.38)	36.01(1.86)
Soft Imput	[0.24(0.02), 1]	**3.76**(0.07)	69.73(3.83)	99.32(0.03)	**48.57**(2.05)	37.28(1.50)
MICE	[.50(0.00),1]	4.46(0.05)	28.57(3.54)	**99.72**(0.07)	32.14(1.75)	**37.50**(3.55)

Table 4: The mean(sd) of model fit and edge detection metrics for NonParanormal Data: $n = 400, p = 300$, MP 2

Method	$\hat{\lambda}^*, \hat{\alpha}^*/\hat{\tau}^*$	$\|\hat{\Omega} - \Omega\|_F$	TPR	TNR	F_1	PPV
DSGL	[0.15(0.00), 1.00(0.00)]	**3.38**(0.04)	**85.30**(1.50)	98.91(0.05)	45.66(1.27)	31.18(1.10)
Thresholded	$\tau = .15\lambda^*$	3.40(0.05)	78.17(2.10)	**99.45**(0.02)	**57.30**(1.60)	**45.24**(1.43)
Comp Only	[0.20(0.00), 1]	3.61(0.04)	76.47(1.57)	99.26(0.04)	50.39(1.39)	37.58(1.33)
Soft Imput	[0.19(0.01), 1]	3.57(0.08)	77.97(3.03)	99.24(0.05)	50.30(1.35)	37.16(1.29)
MICE	[0.28(0.01), 1]	3.90(0.04)	61.95(3.17)	99.38(0.04)	46.06(1.67)	36.69(1.49)

only considering edges detected across all iterations. Next we considered 80% of detected edges and thresholding, choosing τ based on the best performing value from the multivariate normal simulations. We observed stronger intramodality in PET and more intermodality connections between the CSF and MRI modalities. This behavior is still visible in Figure 2b after thresholding.



(a) $\hat{s}_0 = 1$ (b) $\hat{s}_0 \geqslant .8, \tau = .005$

Figure 2: ADNI graphical model of MRI (blue), PET (orange), and CSF (yellow) data, based on percentage each edge was detected, and thresholding. $(\lambda^*, \alpha^*) = (.05, .9)$.

5 DISCUSSION

In this study, we proposed DSGL as a precision matrix estimator that used observed incomplete data without imputation for blockwise missing multimodal data. We demonstrated that DSGL outperforms competing methods including established imputation methods, particularly in terms of graphical structure recovery, especially with high levels of missing blockwise data in both standard Gaussian multivariate normal data and nonparemetric extensions. Our study additionally shows that DSGL theoretically converges using established research on graphical model estimation and missing data. DSGL also successfully produced sparse ADNI graphical model estimations.

Although DSGL shows powerful theoretical and experimental results, there are additional scenarios that need to be considered. First, there are multimodal Missing at Random (MAR) and Missing at not Random (MNAR) simulations proposed by Xue and Qu (2021). However, the MNAR scenarios in this work rely on a response variable, so we will need a similar mechanism while maintaining an unsupervised setting. Secondly, a uniform (sub)-Gaussian distribution can be restrictive for multimodal data. Also, the irrepresentability condition is a necessary condition for theoretical analysis of GLASSO based precision estimation methods. Finally, the post-estimation procedures for edge detection would benefit from hypothesis and inference based methods.

For future work, we will expand DSGL into handling MAR and MNAR settings. Next, we will investigate alternative graphical model estimation methods that are less reliant on the sub-Gaussian and irrepresentability conditions like CLIME and local linear approximation approaches (Breheny and Huang, 2011). In terms of edge evaluation tools, False Discovery Rate hypothesis testing and matrix based hypothesis testing are potential post estimation procedures (Cai et al., 2013; Liu, 2013). Additional considerations include various thresholding schemes controlling inter and intramodal connections and extensions of DSGL to more modern structural analyses seen in hub-based research (Tan et al., 2014; Sánchez Gómez et al., 2025). Finally, in terms of real data applications, we are interested in viewing the ADNI data as a dual-modality problem, analyzing other data sets, and adding literature interpretations of the resulting graphical models.

Acknowledgements

This research was supported by the following:

- National Science Foundation (NSF) DMS Grant 2100729.
- J. Graves's research was supported by the North Carolina A&T Chancellor's Distinguished Fellowship; Y. Liu's research was supported in part by NSF DMS 2625396; E. Zhang's research was supported in part by a Seattle University start-up fund.
- Data collection and sharing for the Alzheimer's Disease Neuroimaging Initiative (ADNI) is funded by the National Institute on Aging (National Institutes of Health Grant U19 AG024904). The grantee organization is the Northern California Institute for Research and Education. In the past, ADNI has also received funding from the National Institute of Biomedical Imaging and Bioengineering, the Canadian Institutes of Health Research, and private sector contributions through the Foundation for the National Institutes of Health (FNIH) including generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.

References

Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008. URL `https://www.jmlr.org/papers/volume9/banerjee08a/banerjee08a.pdf`.

Sumanta Basu, Karl Kumbier, James B Brown, and Bin Yu. Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences*, 115(8):1943–1948, 2018. doi: 10.1073/pnas.1711236115.

Peter J. Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577 – 2604, 2008. doi: 10.1214/08-AOS600.

Patrick Breheny and Jian Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, 5(1):232–253, 2011. doi: 10.1214/10-AOAS388.

T. Tony Cai and Ming Yuan. Adaptive covariance matrix estimation through block thresholding. *The Annals of Statistics*, 40(4):2013–2044, 2012. doi: 10.1214/12-AOS999.

T. Tony Cai, Weidong Liu, and Xi Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106 (494):594–607, 2011. doi: 10.1198/jasa.2011.tm10155.

T. Tony Cai, Weidong Liu, and Yin Xia. Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association*, 108(501):265–277, 2013. doi: 10.1080/01621459.2012.758041.

T. Tony Cai, Zhao Ren, and Harrison H. Zhou. Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10(1):1–59, 2016. doi: 10.1214/15-EJS1081.

David L. Donoho, Iain M. Johnstone, Gérard Kerkyacharian, and Dominique Picard. Wavelet shrinkage: Asymptopia? *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(2):301–369, 1995. doi: 10.1111/j.2517-6161.1995.tb02032.x.

Jianqing Fan, Runze Li, Cun-Hui Zhang, and Hui Zou. *Statistical Foundations of Data Science*. Chapman and Hall/CRC, 2020.

Roger Fan, Byoungwook Jang, Yuekai Sun, and Shuheng Zhou. Precision matrix estimation with noisy and missing data. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2810–2819, 2019. URL `https://proceedings.mlr.press/v89/fan19a.html`.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008. doi: 10.1093/biostatistics/kxm045.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. *glasso: Graphical Lasso: Estimation of Gaussian Graphical Models*, 2019. R package version 1.11.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2nd edition, 2009.

Trevor Hastie, Rahul Mazumder, Jason D. Lee, and Reza Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *Journal of Machine Learning Research*, 16(104):3367–3402, 2015. URL http://jmlr.org/papers/v16/hastie15a.html.

OpenAI Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, et al. GPT-4o system card. Technical report, OpenAI, 2024. URL https://api.semanticscholar.org/CorpusID:273662196.

Haoming Jiang, Xinyu Fei, Han Liu, Kathryn Roeder, John Lafferty, Larry Wasserman, Xingguo Li, and Tuo Zhao. *huge: High-dimensional Undirected Graph Estimation*, 2026. URL https://CRAN.R-project.org/package=huge. R package version 1.4.

Mladen Kolar and Eric P Xing. Estimating sparse precision matrices from fata with missing values. In John Langford and Joelle Pineau, editors, *Proceedings of the 29th on International Conference on Machine Learning*, volume 29 of *ICML'12*, pages 635–642, 2012. doi: https://dl.acm.org/doi/10.5555/3042573.3042657.

Mladen Kolar, Han Liu, and Eric P. Xing. Graph estimation from multi-attribute data. *Journal of Machine Learning Research*, 15(51):1713–1750, 2014. URL http://jmlr.org/papers/v15/kolar14a.html.

Mike Laszkiewicz, Asja Fischer, and Johannes Lederer. Thresholded adaptive validation: Tuning the graphical lasso for graph recovery. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1864–1872, 2021. URL https://proceedings.mlr.press/v130/laszkiewicz21a.html.

Hongzhe Li and Jiang Gui. Gradient directed regularization for sparse gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics*, 7(2):302–317, 2006. doi: 10.1093/biostatistics/kxj008.

Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(80):2295–2328, 2009. URL http://jmlr.org/papers/v10/liu09a.html.

Han Liu, Kathryn Roeder, and Larry Wasserman. Stability approach to regularization selection (stars) for high dimensional graphical models. *Advances in Neural Information Processing Systems*, 23(2):1432–1440, 2010. URL https://pubmed.ncbi.nlm.nih.gov/25152607/. PMID: 25152607.

Han Liu, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman. High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293 – 2326, 2012. doi: 10.1214/12-AOS1037.

Weidong Liu. Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics*, 41(6):2948–2978, 2013. doi: 10.1214/13-AOS1169.

Po-Ling Loh and Xin Lu Tan. High-dimensional robust precision matrix estimation: Cellwise corruption under ϵ-contamination. *Electronic Journal of Statistics*, 12(1):1429–1467, 2018. doi: 10.1214/18-EJS1427.

Rahul Mazumder and Trevor Hastie. Exact covariance thresholding into connected components for large-scale graphical lasso. *Journal of Machine Learning Research*, 13:781–794, 2012. URL http://jmlr.org/papers/v13/mazumder12a.html.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11(80):2287–2322, 2010. URL http://jmlr.org/papers/v11/mazumder10a.html.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006. doi: 10.1214/009053606000000281.

Susanne G. Mueller, Michael W. Weiner, Leon J. Thal, Ronald C. Petersen, Clifford Jack, William Jagust, John Q. Trojanowski, Arthur W. Toga, and Laurel Beckett. The alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics of North America*, 15(4):869–877, xi–xii, 2005. doi: 10.1016/j.nic.2005.09.008.

Seongoh Park, Xinlei Wang, and Johan Lim. Estimating high-dimensional covariance and precision matrices under general missing dependence. *Electronic Journal of Statistics*, 15(2):5301–5340, 2021. doi: 10.1214/21-EJS1892.

Pradeep Ravikumar, Martin J. Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011. doi: 10.1214/11-EJS631.

José Á. Sánchez Gómez, Weibin Mo, Junlong Zhao, and Yufeng Liu. Hub detection in gaussian graphical models. *Journal of the American Statistical Association*, 120(552):2397–2409, 2025. doi: 10.1080/01621459.2025.2453250.

Somayeh Sojoudi. Equivalence of graphical lasso and thresholding for sparse graphs. *Journal of Machine Learning Research*, 17(115):1–21, 2016. URL http://jmlr.org/papers/v17/16-013.html.

Shanshan Song, Yuanyuan Lin, and Yong Zhou. Semi-supervised inference for block-wise missing data without imputation. *Journal of Machine Learning Research*, 25 (99):1–36, 2024. URL https://www.jmlr.org/papers/v25/21-1504.html.

Nicolas Städler and Peter Bühlmann. Missing values: Sparse inverse covariance estimation and an extension to sparse regression. *Statistics and Computing*, 22(1):219–235, 2012. doi: 10.1007/s11222-010-9219-7.

Sandra Steyaert, Marija Pizurica, Divya Nagaraj, Priya Khandelwal, Tina Hernandez-Boussard, Andrew J. Gentles, and Olivier Gevaert. Multimodal data fusion for cancer biomarker discovery with deep learning. *Nature Machine Intelligence*, 5(4):351–362, 2023. doi: 10.1038/s42256-023-00633-5.

Kean Ming Tan and Daniela Witten. Statistical properties of convex clustering. *Electronic Journal of Statistics*, 9 (2):2324–2347, 2015. doi: 10.1214/15-EJS1074.

Kean Ming Tan, Palma London, Karthik Mohan, Su-In Lee, Maryam Fazel, and Daniela Witten. Learning graphical models with hubs. *Journal of Machine Learning Research*, 15(95):3297–3331, 2014. URL http://jmlr.org/papers/v15/tan14b.html.

Katherine Tsai, Boxin Zhao, Sanmi Koyejo, and Mladen Kolar. Latent multimodal functional graphical model estimation. *Journal of the American Statistical Association*, 119(547):2217–2229, 2024. doi: 10.1080/01621459.2023.2252142.

Jitendra K Tugnait. Sparse-group lasso for graph learning from multi-attribute data. *IEEE Transactions on Signal Processing*, 69:1771–1786, 2021. doi: 10.1109/TSP.2021.3057699.

Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011. doi: 10.18637/jss.v045.i03.

Janani Venugopalan, Li Tong, Hamid Reza Hassanzadeh, and May D. Wang. Multimodal deep learning models for early detection of alzheimer's disease stage. *Scientific Reports*, 11(3254), 2021. doi: 10.1038/s41598-020-74399-w.

Huahua Wang, Farideh Fazayeli, Soumyadeep Chatterjee, and Arindam Banerjee. Gaussian copula precision estimation with missing values. In Samuel Kaski and Jukka Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 978–986, 2014. URL https://proceedings.mlr.press/v33/wang14a.html.

Shuo Xiang, Lei Yuan, Wei Fan, Yalin Wang, Paul M Thompson, Jieping Ye, Alzheimer's Disease Neuroimaging Initiative, et al. Bi-level multi-source learning for heterogeneous block-wise missing data. *NeuroImage*, 102(1):192–206, 2014. doi: 10.1016/j.neuroimage.2013.08.015.

Fei Xue and Annie Qu. Integrating multisource block-wise missing data in model selection. *Journal of the American Statistical Association*, 116(536):1914–1927, 2021. doi: 10.1080/01621459.2020.1751176.

Guan Yu, Quefeng Li, Dinggang Shen, and Yufeng Liu. Optimal sparse linear prediction for block-missing multi-modality data without imputation. *J. Am. Stat. Assoc.*, 115(531):1406–1419, 2020. doi: 10.1080/01621459.2019.1632079. PMID: 34824484.

Lei Yuan, Yalin Wang, Paul M. Thompson, Vaibhav A. Narayan, and Jieping Ye. Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. *NeuroImage*, 61(3):622–632, 2012. doi: https://doi.org/10.1016/j.neuroimage.2012.03.059.

Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007. doi: 10.1093/biomet/asm018.

Fei Zhao, Chengcui Zhang, and Baocheng Geng. Deep multimodal data fusion. *ACM computing surveys*, 56(9): 1–36, 2024. doi: 10.1145/3649447.

Tuo Zhao, Han Liu, Kathryn Roeder, John Lafferty, and Larry Wasserman. The huge package for high-dimensional undirected graph estimation in r. *Journal of Machine Learning Research*, 13(37):1059–1062, 2012. URL http://jmlr.org/papers/v13/zhao12a.html.

Shuheng Zhou, Philipp Rütimann, Min Xu, and Peter Bühlmann. High-dimensional covariance estimation based on gaussian graphical models. *Journal of Machine Learning Research*, 12(91):2975–3026, 2011. URL http://jmlr.org/papers/v12/zhou11a.html.

Gaussian Graphical Learning via PSD Constraint for Blockwise Missing Multimodal Data (Supplementary Material)

Joseph Graves[1] **Yufeng Liu**[2] **Elio Zhang**[3] **Seong-Tae Kim**[1]

for the Alzheimer's Disease Neuroimaging Initiative*

[1]Mathematics and Statistics Department, North Carolina A & T State University, Greensboro, NC, USA
[2]Department of Statistics, University of Michigan, Ann Arbor, MI, USA
[3]Mathematics Department, Seattle University, Seattle, WA, USA

A THEORETICAL SECTION: PROOFS

In order to justify the convergence of DSGL's precision estimator, we use the body of work established by Ravikumar et al. (2011). The theoretical work from this paper established the existence of a unique solution to the log-likelihood optimization problem, and established the conditions necessary for the elementwise and graphical convergence of an estimator $\hat{\Omega}$ to the true precision matrix Ω. We also build on the extensions of Ravikumar et al. (2011) that focuses on the context of missing data, which addressed the MCAR and multimodal settings (Kolar and Xing, 2012; Yu et al., 2020).

A.1 TAIL CONDITIONS

The first step is to ensure we have a reasonable bound on the elementwise difference of our pilot estimator $\hat{\Sigma}_{\nu,\alpha}$ from Σ. This is known as a tail condition defined below:

Definition A.1 (Tail Condition)**.** *Let X be a random vector. Then X satisfies tail condition $\mathcal{T}(f, M)$ if, for some constant M and positive, monotonic, real valued function $f(n, \delta)$, $n \in \mathbf{N}$, $\delta > 0$:*

$$P(|\hat{\Sigma}_{ij} - \Sigma_{ij}| \geqslant \delta) \leqslant 1/f(n, \delta) \text{ for all } \delta < 1/M.$$

Additionally, the inverse tail functions are defined as the following:

Definition A.2 (Inverse Tail Functions)**.** *Let $f(n, \delta)$ be a positive, monotonic, real valued function associated with tail condition $\mathcal{T}(f, M)$, then the inverse tail functions are:*

$$\bar{n}(\delta, r) = \arg\max\{n | f(n, \delta) \leqslant r\}$$

and

$$\bar{\delta}(n, r) = \arg\max\{\delta | f(\delta, n) \leqslant r\}.$$

By definition of the inverse tail functions and the property of monotonicity, we have the following property:

Lemma A.3. *If $n > \bar{n}(\delta, r)$ for some δ, then $\delta \geqslant \bar{\delta}(n, r)$.*

Proof by contradiction: Suppose $\delta < \bar{\delta}(n, r)$, then $f(n, \delta) \leqslant r \Rightarrow n \leqslant \bar{n}(\delta, r)$.

As we are not using the standard empirical sample covariance matrix, we must ensure that $\hat{\Sigma}_{\nu,\alpha}$ satisfies a set tail condition, as this is the preceding step to proving Theorems 3.2 and 3.3.

We first begin by stating the conditions for a standard empirical covariance estimator to satisfy a specified tail bound.

Lemma A.4 (Lemma 1 of Ravikumar et al. (2011))**.** *Let $(X_1, \ldots, X_p)$ be a zero-mean random vector with true covariance Σ so that $X_i/\sqrt{\Sigma_{ii}}$ is sub-Gaussian with parameter L. Then for sample covariance estimator $\hat{\Sigma}$, we have*

$$P\left(|\hat{\Sigma}_{ij} - \Sigma_{ij}| \geqslant \delta\right) \leqslant 4\exp\left\{-\frac{n\delta^2}{128(1+4L^2)^2 \max_i(\Sigma_{ii})^2}\right\}$$

for all δ in $\left(0, 8(1+4L^2)\max_i(\Sigma_{ii}\right)$.

Subsequent works analyzing missing data develop Lemma A.4 to ensure their proposed pilot estimator satisfies such a tail bound (Kolar and Xing, 2012; Yu et al., 2020). In particular, both cited works show that an unbiased pilot estimator for Σ can satisfy Lemma A.4 as long as the sub-Gaussian conditions in Assumption 3.1 and for $X_i/\sqrt{\Sigma_{ii}}$ hold. We approach the proof of Theorem 3.1 with this strategy. To simplify our proofs, we add the following assumption:

Assumption A.1 (Scaling of $\tilde{\Sigma}$)**.** *Let $N_i = |S_i|$ where $S_i = \{j : x_{ji}$ is observed $\}$. Then assume X_{ij} is scaled such that $\sum_{j \in S_i} x_{ji}^2 = N_i$. Therefore $\tilde{\Sigma}_{ii} = 1$ for all i.*

Proof of Theorem 3.1:

We first ensure the tail bound on the transformed preliminary estimator, $\tilde{\Sigma}_Z$. We satisfy the sub-Guassian assumption for $\tilde{\Sigma}_Z/(\Sigma_Z)_{ii}$ per Assumption 3.1. By Theorem 1 of Yu et al. (2020), we are able to conclude that $\tilde{\Sigma}_Z$ satisfies a specified tail condition. However, we proceed to define $\hat{\Sigma}_{\nu,\alpha} = \frac{1}{1+\nu}\tilde{\Sigma}_D + \frac{\nu}{1+\nu}I_p + \frac{\alpha}{1+\nu}\tilde{\Sigma}_C$. Immediately have each coefficient is less than 1 as $\nu > 0$ and $0 < \alpha \leqslant 1$.

Then Assumption 3.2 implies that $\delta^* \leqslant \dfrac{8}{\sqrt{1+\nu}}(1+4L^2)$, allowing the use of Lemma A.4 to state:

$$P\left(|(\tilde{\Sigma}_Z - \Sigma_Z)_{ij}| \geqslant \delta\right) \leqslant 4\exp\left\{-\frac{n\delta^2}{128(1+4L^2)^2 \max_i(\Sigma_{ii})^2}\right\}.$$

We then substitute δ with δ^* and use $\mathrm{diag}(\Sigma_Z) = 1/(1+\nu)$ to reach the elementwise bound of δ^*. We reach the probability bound of $4p^{-3\varphi}$ for $\tilde{\Sigma}_Z$ by using Assumption 3.2 once more:

$$P\left(|(\tilde{\Sigma}_Z - \Sigma_Z)_{ij}| \geqslant \delta^*\right) \leqslant 4p^{-3\varphi}. \tag{11}$$

Let $\hat{\Sigma}_{\nu,\alpha}$ be defined as in Algorithm 1. We then consider $(\hat{\Sigma}_{\nu,\alpha})_{ij} - (\Sigma_Z)_{ij}$ for $i = j$, $i \neq j$ with $M_i = M_j$, and $i \neq j$ for $M_i \neq M_j$.

It follows that, by Assumption 2.2:

$$\left(\hat{\Sigma}_{\nu,\alpha} - \Sigma_Z\right)_{ij} = \begin{cases} \frac{\nu}{1+\nu} & i = j \\ \frac{1}{1+\nu}(\tilde{\sigma}_{ij} - \sigma_{ij}) & i \neq j, \text{ same modality} \\ \frac{1}{1+\nu}(\alpha\tilde{\sigma}_{ij} - \sigma_{ij}) & i \neq j, \text{ different modality.} \end{cases}$$

For $i \neq j$ within the same modality, we have $|\tilde{\sigma}_{ij} - \sigma_{ij}|$ which matches Equation (11).

For $i \neq j$ for differing modalities, we have $\frac{1}{1+\nu}|\alpha\tilde{\sigma}_{ij} - \alpha\sigma_{ij} - (1-\alpha)\sigma_{ij}| \leqslant \frac{\alpha}{1+\nu}|\tilde{\sigma}_{ij} - \sigma_{ij}| + \frac{1-\alpha}{1+\nu} \leqslant \alpha\delta^* + \frac{1-\alpha}{1+\nu}$. In order to use the A.4, we must ensure this quantity is within the domain of the tail function.

We consider two possibilities. If $\delta^* > \frac{1}{1+\nu}$, then $\alpha\delta^* + \frac{1-\alpha}{1+\nu} < \alpha\delta^* + \delta^*(1-\alpha) = \delta^*$, which we know is within the domain. If $\delta^* \leqslant \frac{1}{1+\nu}$, then $\alpha\delta^* + \frac{1-\alpha}{1+\nu} \leqslant \frac{1}{1+\nu} \leqslant \frac{1}{\sqrt{1+\nu}} \leqslant \frac{1}{\sqrt{1+\nu}}8(1+4L^2)$.

Therefore we conclude that the elementwise bound of $\hat{\Sigma}_{\nu,\alpha}$ is restricted by the stated quantity with probability $1 - 4p^{-3\varphi}$.

Remark: In addition to the remark made at the end of Theorem 3.1, we are able to state the satisfied tail condition $T(f, v_*)$, with $v_* = [8(1 + 4L^2)]^{-1}\sqrt{1+\nu}$. With the sub-Gaussian condition, we define an explicit expression for the the inverse tail functions.

$$\bar{\delta}(r; n) = \sqrt{\frac{\log(4r)}{c_* n}} \text{ and } \bar{n}(r; \delta) = \frac{\log(4r)}{c_* \delta^2}.$$

Corollary 3.1.1 therefore states that the inverse tail function $\bar{\delta}$ evaluated at (n, p^{φ^*}) is equal to δ^*, which is an important quantity for Theorem 3.2.

A.2 THEOREM 3.2

Theorem 3.2 regarding the elementwise convergence of the DSMG estimator is motivated by Theorem 1 of Ravikumar et al. (2011). For our proof strategy, we follow the series of Lemmas developed by Ravikumar et al. (2011) to prove that the solution to GLASSO $\hat{\Omega}$ converges to the true Ω. We define the necessary terminology. First, a core component of the theoretical work of Ravikumar et al. (2011) is the primal-dual witness approach, where the primal estimator for Ω is defined such that

$$\tilde{\Omega} := \underset{\Theta>0, \Theta=\Theta^T, \Theta_{S^C}=0}{\arg\min} - \log\det(\Theta) + \text{tr}(\hat{\Sigma}_{\nu,\alpha}\Theta) + \lambda\|\Theta\|_{1,\text{off}}.$$

$\tilde{\Omega}$, the solution to GLASSO constrained to the true-edge set is a purely theoretical quantity. Under certain conditions, then $\tilde{\Omega}$ serves as the witness to the non-restricted GLASSO problem. For simplicity, define $\hat{\Sigma} = \hat{\Sigma}_{\nu,\alpha}, \hat{\Omega} = \hat{\Omega}(\nu, \alpha)$.

$$W := \hat{\Sigma} - \Sigma$$
$$\Delta := \tilde{\Omega} - \Omega$$
$$R(\Delta) := \tilde{\Omega}^{-1} - \Omega^{-1} + \Omega^{-1}\Delta\Omega^{-1}.$$

We then state the following Lemmas:

Lemma A.5 (Lemma 8 of Ravikumar et al. (2011)). *Let $\varphi > 2$ and n be the sample size such that $\bar{\delta}(n, p^{\varphi}) \leqslant 1/v^*$. Then:*

$$P\left[\|W\|_\infty \geqslant 1/v^*\right] \leqslant \frac{1}{p^{\varphi-2}}.$$

Lemma A.6 (Lemmas 3 and 4 of Ravikumar et al. (2011)). *If $\max\{\|W\|_\infty, \|R(\Delta)\|_\infty\} \leqslant \gamma\lambda/8$, then the restrained GLASSO estimate $\tilde{\Omega}$ satisfies the necessary conditions of the dual primal-witness approach such that $\tilde{\Omega} = \hat{\Omega}$.*

Lemma A.7 (Lemma 5 of Ravikumar et al. (2011)). *Let $\|\Delta\|_\infty \leqslant (3\kappa_\Sigma d)^{-1}$. Then $\|R(\Delta)\|_\infty \leqslant \frac{3}{2}d\|\Delta\|_\infty^2\kappa_\Sigma^3$.*

Lemma A.8 (Lemma 6 of Ravikumar et al. (2011)). *If $r := 2\kappa_\Gamma(\|W\|_\infty + \lambda) \leqslant \min\left\{\frac{1}{3\kappa_\Sigma d}, \frac{1}{3\kappa_\Sigma^3\kappa_\Gamma d}\right\}$.*

Then $\|\Delta\|_\infty \leqslant r$.

Our goal therefore, is to show that $\hat{\Omega}$ and its counterpart $\tilde{\Omega}$ can be constructed and satisfy the necessary properties to conclude the elementwise convergence of $\hat{\Omega}$ to Ω. We take particular precautions as we are using a pilot estimator built from blockwise missing multimodal data rather than the sample covariance. We define $\lambda^* := \left(\frac{8}{\gamma}\right)\delta^*$ which follows from Corollary 3.1.1, allowing for us to begin our proof.

Proof of Theorem 3.2

Define Event $\mathcal{A}$:

$$\mathcal{A} = \|W\|_\infty \leqslant \max\left\{A'\sqrt{\frac{\log(p^\varphi)}{tn}}, \frac{\nu}{1+\nu}\right\}. \tag{12}$$

Then by Theorem 3.1 and Lemma A.5, we can use union probability to conclude that

$$P(\mathcal{A}) \geqslant 1 - 4p^{-3\varphi+2}.$$

Subsequent analysis will therefore be within Event $\mathcal{A}$. We define $\lambda^* = (8/\gamma)\delta^* = (8/\gamma)A'\sqrt{\frac{\log(p^{\varphi^*})}{tn}}$. Our main goal is to prove Lemma A.6, so we must ensure both conditions are satisfied. Immediately, by Equation (12) and the definition of λ^*, we have $\|W\|_\infty \leqslant \gamma\lambda^*/8$. It remains to show the other condition is satisfied using Lemmas A.8 and A.7.

Evaluating r, we have $r = 2\kappa_\Gamma(\|W\|_\infty + \lambda^*) \leqslant (1 + 8/\gamma)\delta^*$. By Corollary 3.1.1, we have that $\delta^* = \bar{\delta}(n, p^{\varphi^*})$, which allows us to use the Lemma A.3 and the formal conditions of Theorem 1 of Ravikumar et al. (2011) to conclude:

$$r \leqslant \frac{2\kappa_\Gamma(1+8/\gamma)}{6(1+8/\gamma)d\max\{\kappa_\Sigma\kappa_\Gamma, \kappa_\Sigma^3\kappa_\Gamma^2\}} \leqslant \min\left\{\frac{1}{3\kappa_\Sigma d}, \frac{1}{3d\kappa_\Sigma^3\kappa_\Gamma}\right\}. \tag{13}$$

Therefore, we can now use Lemma A.8 to state:

$$\|\Delta\|_\infty \leqslant 2\kappa_\Gamma(\|W\|_\infty + \lambda^*) \leqslant 2\kappa_\Gamma(1+8/\gamma)\delta^*. \tag{14}$$

Equations (13, 14) immediately satisfies the conditions for Lemma A.7. Finally, we focus on Lemma A.6:

$$\begin{aligned}
\|R(\Delta)\|_\infty &\leqslant \frac{3}{2}d\|\Delta\|_\infty^2\kappa_\Sigma^3 \\
&\leqslant 6\kappa_\Sigma^3\kappa_\Gamma^2 d(1+8/\gamma)^2[\delta^*]^2 \\
&\leqslant [6\kappa_\Sigma^3\kappa_\Gamma^2 d(1+8/\gamma)^2\delta^*]\frac{\gamma\lambda^*}{8} \\
&\leqslant \alpha\lambda^*/8.
\end{aligned}$$

The final line comes from the condition of Theorem 3.2

$$48\kappa_\Sigma^3\kappa_\Gamma^2(1+\gamma/8)^2 d\lambda^* \leqslant \gamma$$

introduced by Loh and Tan (2018). With the conditions for Lemma A.6 satisfied, we can conclude that $\tilde{\Omega} = \hat{\Omega}$ and that their respective edge-set complements align, i.e., the second conclusion to Theorem 3.2. Furthermore, we can state $\Delta = \tilde{\Omega} - \Omega = \hat{\Omega} - \Omega$, which allows us to state the elementwise convergence of $\hat{\Omega}$ seen in Theorem 3.2.

A.3 THEOREM 3.3

Proof of Theorem 3.3

The proof of Theorem 3.3 uses similar tactics using Lemma A.3 to satisfy the following conditions for the following Lemma.

Lemma A.9 (Lemma 7 of Ravikumar et al. (2011))**.** *Let the conditions for Lemma A.8 hold and that ω_{min} is lower bounded s.t.*

$$\omega_{min} \geqslant 4(\|W\|_\infty + \lambda).$$

Then the sign of $\tilde{\Omega}$ matches the sign of Ω.

Using the results of Theorem 3.2, it follows, under Event $\mathcal{A}$, that $\|\tilde{\Omega} - \Omega\|_\infty < \omega_{min}/2$, allowing the use of Lemma A.9 to conclude the edge set of Ω matches the edge set of $\hat{\Omega}$.

B ADDITIONAL SIMULATION RESULTS

We report the following supplementary simulation results: the $p = 30$ simulation results for both multivariate normal and nonparanormal results. We also show the failure rate of DISCOM vs DSGL.

B.1 MULTIVARIATE NORMAL, P = 30

We generate the true precision matrix Ω_{30} with 10 predictors per modality using the same random graph generation rule as in $p = 300$. The notable difference is that we demonstrate an alternative rule for the maximum candidate λ-value: $\sqrt{\log(p)/tn}$. This is a useful limit when there is concern that DSGL chooses too high of a λ value leading to zero edges being detected.

Table 5: The mean(sd) of model fit and edge detection metrics for Multivariate Normal Random Graph, $n = 400, p = 30$, MP 1.

Method	$\hat{\lambda}^*, \hat{\alpha}^*/\hat{\tau}^*$	$\|\hat{\Omega} - \Omega\|_F$	TPR	TNR	F_1	PPV
DSGL	[0.17(0.01), .94(0.07)]	1.06(0.07)	99.58(1.59)	99.40(0.47)	92.81(4.84)	87.31 (8.29)
DSGL, thresh	$\tau = .25\lambda^*$	1.06(0.07)	97.92(3.79)	99.98(0.07)	98.61(2.20)	99.40 (1.83)
MICE	[0.22(0.03), 1]	1.19(0.13)	99.88(5.86)	99.88(0.21)	96.84(3.57)	97.17 (4.56)
Comp Only	[0.25 (0.02), 1]	1.31(0.09)	95.62(4.69)	99.95(0.12)	97.11(2.72)	98.81(2.83)
Soft Imput	[0.46(0.06), 1]	1.73(0.07)	27.50(23.88)	100.00(0.00)	38.28(27.22)	100.00(0.00)

Table 6: The mean(sd) of model fit and edge detection metrics for Multivariate Normal Random Graph, $n = 400, p = 30$, MP2

Method	$\hat{\lambda}^*, \hat{\alpha}^*/\hat{\tau}^*$	$\|\hat{\Omega} - \Omega\|_F$	TPR	TNR	F_1	PPV
DSGL	[0.12(.00), 1.00(.00)]	0.82(.06)	100.00(.00)	97.28(.80)	74.15(5.64)	59.23(7.14)
DSGL, thresh	$\tau = .45\lambda^*$	0.82 (.07)	99.17(2.16)	99.99(0.04)	99.47(1.21)	99.80(1.07)
MICE	[0.13(.02), 1]	0.84(.09)	100.00(.00)	97.02(1.99)	74.00(11.77)	60.03(14.38)
Comp Only	[0.12(0.00), 1]	0.82(0.06)	100.00(.00)	97.28(0.80)	74.15(5.64)	59.23(7.14)
Soft Imput	[0.27(0.01), 1]	1.38(0.08)	93.12(4.74)	99.98(0.07)	96.08(2.65)	99.37(1.91)

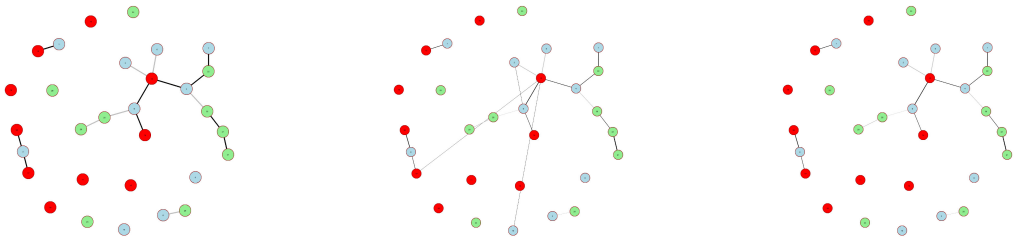
We note similar trends as seen in the main manuscript. In particular, DSGL consistenly outperforms competing methods in terms of TPR and its standard deviation. A higher PPV is recovered upon introducing thresholding without sacrificing TPR and an improved consistency. Ultimately, we demonstrate DSGL performs well or better than competing methods in both low and high dimensional cases, where more instability is present. Figures 3 and 4 show a precision matrix estimate from each method contrasted to the true Ω for $p = 30$, color coded by modality. All methods in the first missing pattern capture the key structure, with the notable exception of soft imputation, as seen in Table 5 in terms of the high TPR standard deviation. In general, DSGL recovers the true graphical structure quite well upon introducing thresholding.

B.2 FAILURE RATE OF DISCOM

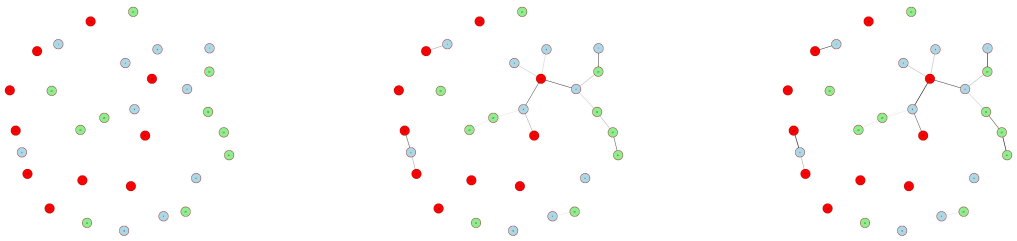
In the original DISCOM method by Yu et al. (2020), the pilot covariance estimator trains coefficients for $\tilde{\Sigma}_D$ in addition to $\tilde{\Sigma}_C$ s.t.

$$\hat{\Sigma}_{DISCOM} = \alpha_1 \tilde{\Sigma}_D + \alpha_2 \tilde{\Sigma}_C + (1 - \alpha_1)\frac{\text{tr}(\tilde{\Sigma})}{p}.$$

Note that $\alpha_i \in (0, 1]$. However, in the simulation code, combinations of α_1, α_2 that yield a non-positive definite $\hat{\Sigma}$ are ignored. In contrast, our proposed method DSGL makes $\hat{\Sigma}$ positive definite using the minimum eigenvalue. We therefore, compare the failure rate of DISCOM compared to DSGL while also reporting the parameter selection of DISCOM. Tables 7, 8 show results for $p = 30$ and $p = 300$. The failure rate of DISCOM is significantly higher in the higher-dimensional scenario. Furthermore, DSGL's α parameter is larger than the corresponding α_2 parameter of DISCOM while converging to 1. Interestingly, we see inverse effects on the choice of λ. In $p = 30$ the optimal λ decreased with more observable data available, but in $p = 300$, the optimal λ increases in magnitude. We note that DISCOM can also achieve positive-definiteness using the minimum eigenvalue as seen in DSGL.

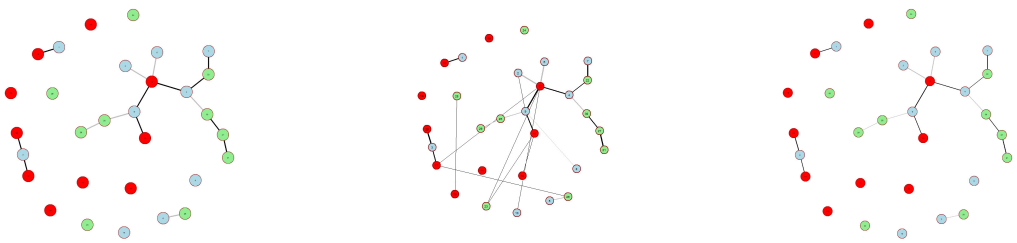


(a) True Ω (b) DSGL (c) DSGL,thresholded

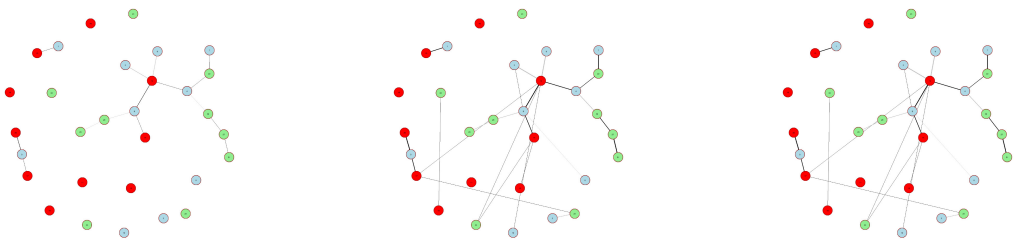


(d) Soft Imputation (e) Complete Only (f) MICE

Figure 3: Graphical Model Comparison for Multivariate Normal Missing Pattern 1



(a) True Ω (b) DSGL (c) DSGL,thresholded



(d) Soft Imputation (e) Complete Only (f) MICE

Figure 4: Graphical Model Comparison for Multivariate Normal Missing Pattern 2

Table 7: DSGL vs DISCOM, Multivariate Normal, n = 400, p = 30

MP	Method	$\hat{\lambda}$	$\hat{\alpha}$	Failure Rate(%)
1	DSGL	0.17(0.01)	α =**0.94**(0.07)	0
1	DISCOM	0.11(0.01)	α_1 =0.98(0.03), α_2 =0.97(0.04)	16.48(4.75)
2	DSGL	0.12(0.00)	α =**1**(0.00)	0
2	DISCOM	0.10(0.00)	α_1=1(0), α_2=1.00(0.02)	0(0)

Table 8: DSGL vs DISCOM, Multivariate Normal, n = 400, p = 300

MP	Method	$\hat{\lambda}$	$\hat{\alpha}$	Failure Rate(%)
1	DSGL	0.13(0.01)	α =**0.65**(0.01)	0
1	DISCOM	0.10(0.00)	α_1 =0.61(0.06), α_2 =0.39(0.02)	70.45(0.32)
2	DSGL	0.15(0.00)	α =**1**(0.00)	0
2	DISCOM	0.12(0.00)	α_1=0.76(0.02), α_2=0.75(0.00)	46.49(0.36)

B.3 NONPARANORMAL, P = 30

We construct the semiparametric Gaussian copula estimation of the nonparanormal distribution for $p = 30$ using the same function transformations of the original multivariate Gaussian data. Table 9 shows the Nonparanormal results for Missing Pattern 1. In Missing Pattern 1, DSGL achieved the lowest Frobenius norm of all methods and the highest True Positive Rate. Soft Imputation only detected two percent of the total positive edges on average. DSGL's closest competitors were MICE and the Complete Only case, having higher Positive Predictive Value scores with lower stanard deviation than DSGL, but had significantly lower True Positive Rates. The thresholded version of DSGL using the stated rule in Equation 10 improved the Positive Predictive Value significantly in mean and standard deviation with a small loss in the True Positive Rate in the same fields. However, DSGL maintained the highest F_1 score and had the lowest standard deviation of all the compared methods. In Table 10, we observe that MICE and Complete Only both improve and match base DSGL's metrics, with notably worse standard deviations in all fields. DSGL's optimal α^* approaches 1 with more observable data, and the three aforementioned methods see worsening Positive Predictive Values. This drop from Missing Pattern 1 to Missing Pattern 2 occurred in the Multivariate Normal simulation example as well for $p = 30$. However, thresholding significantly improves DSGL's Positive Predictive Value in both mean and standard deviation while keeping its True Positive Rate above 90 percent. Overall, DSGL remained robust under thresholding in this scenario, showing it performs well in the Nonparanormal Setting in a low-dimension. Soft Imputation recovers, but not as well as the competing methods. Figures 5 and 6 display visual representations of the graphical estimates.

C SIMULATION CODE

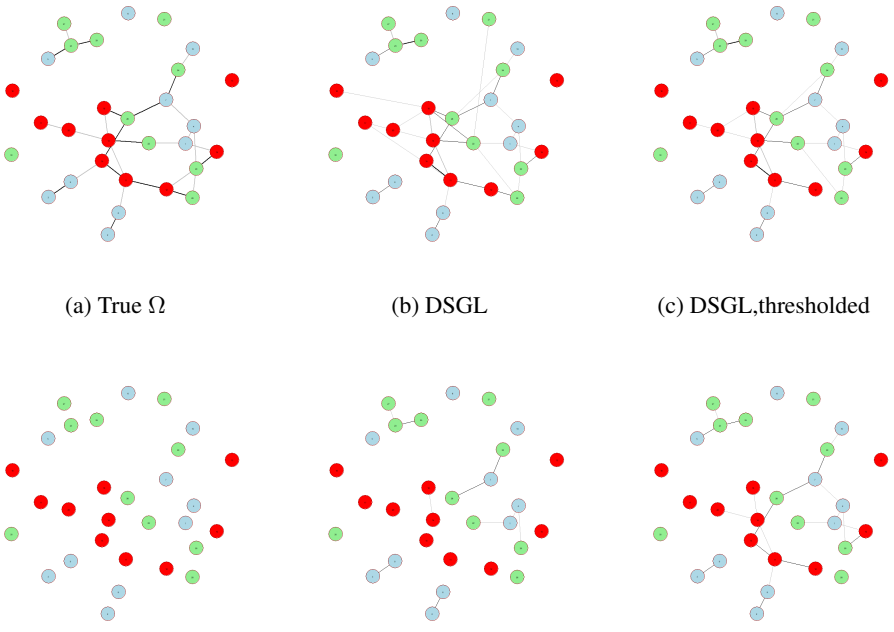
All corresponding code will be provided in a zip file and be available in the following Github, `https://github.com/jgraves50/UAI_DSGL`. The files contain all defined functions organized by data generation, estimation, and algorithms. The notable exception is the ADNI data, which cannot be shared. We provide a general sample code with similar dimension size to run the method.

Table 9: The mean(sd) of model fit and edge detection metrics for Nonparanormal Data, $n = 400, p = 30$, Missing Pattern (MP) 1.

Method	$\hat{\lambda}^*, \hat{\alpha}^*/\hat{\tau}^*$	$\|\hat{\Omega} - \Omega\|_F$	TPR	TNR	F_1	PPV
DSGL	[0.12(0.01), .70(0.07)]	**1.30**(0.07)	**91.90**(5.30)	98.55(0.94)	86.56(5.51)	82.48 (9.23)
DSGL, thresh	$\tau = .1\lambda^*$	1.39(0.07)	87.14(7.42)	99.43(0.47)	**89.15**(4.50)	91.94(6.24)
MICE	[0.23(0.06), 1]	1.61(0.18)	69.76(23.25)	99.93(0.14)	78.83(22.05)	98.97 (2.19)
Comp Only	[0.23 (0.02), 1]	1.65(0.10)	68.81(12.05)	99.99(.04)	80.84(9.06)	99.86(0.76)
Soft Imput	[0.47(0.05), 1]	1.98(0.01)	2.02(4.37)	**100.00**(0.00)	3.65(7.70)	**100.00**(0.00)

Table 10: The mean(sd) of model fit and edge detection metrics for Nonparanormal Data, $n = 400, p = 30$, MP 2.

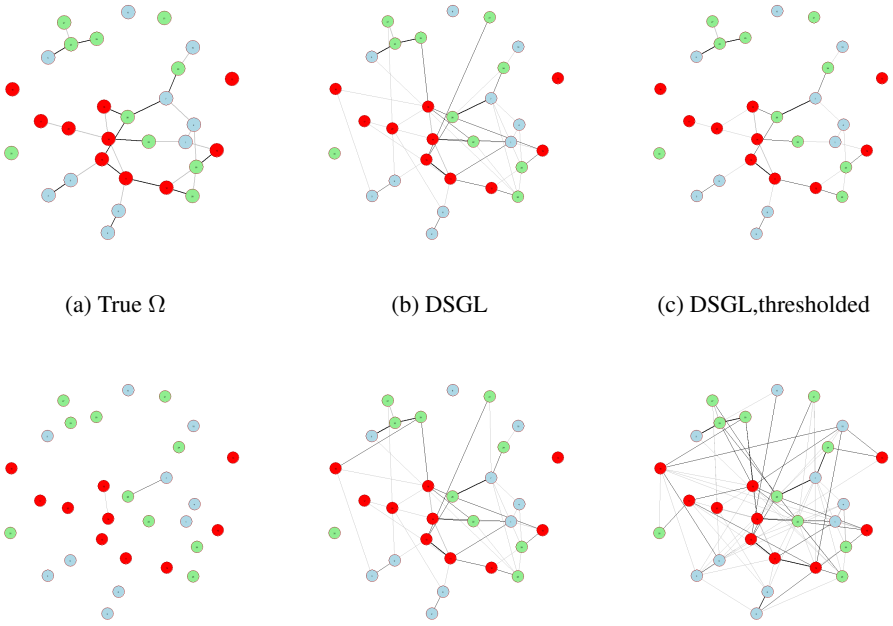
Method	$\hat{\lambda}^*, \hat{\alpha}^*/\hat{\tau}^*$	$\|\hat{\Omega} - \Omega\|_F$	TPR	TNR	F_1	PPV
DSGL	[0.10(0.00), .96(0.03)]	0.99(0.06)	98.57(2.01)	94.94(1.06)	72.66(4.36)	57.70(5.33)
DSGL, thresh	$\tau = .5\lambda^*$	1.00(0.07)	92.86(4.78)	99.81(0.18)	**94.90**(2.73)	97.22(2.61)
MICE	[0.10(0.02), 1]	0.98(0.12)	98.33(2.61)	94.12(4.06)	71.78(12.42)	58.22(15.85)
Comp Only	[0.10(0.00), 1]	**0.96**(0.06)	**98.93**(1.91)	94.55(1.27)	71.44(5.35)	56.16(6.79)
Soft Imput	[0.29(0.06), 1]	1.84(0.08)	41.07(18.11)	**100.00**(0.00)	55.62(21.30)	**100.00**(0.00)



(a) True Ω (b) DSGL (c) DSGL,thresholded

(d) Soft Imputation (e) Complete Only (f) MICE

Figure 5: Graphical Model Comparison for Nonparanormal Missing Pattern 1



(a) True Ω (b) DSGL (c) DSGL,thresholded

(d) Soft Imputation (e) Complete Only (f) MICE

Figure 6: Graphical Model Comparison for Nonparanormal Missing Pattern 2