Robust Multi-view Co-expression Network Inference

Teodora Pandeva
University of Amsterdam

Martijs Jonker
University of Amsterdam

Leendert Hamoen
University of Amsterdam

Joris Mooij
University of Amsterdam

Patrick Forré
University of Amsterdam

Abstract

Unraveling the co-expression of genes across studies enhances the understanding of cellular processes. Inferring gene co-expression networks from transcriptome data presents many challenges, including the high-dimensionality of the data relative to the number of samples, sample correlations, and batch effects. To address these complexities, we introduce a robust method for high-dimensional graph inference from multiple independent studies. We base our approach on the premise that each dataset is essentially a noisy linear mixture of gene loadings that follow a multivariate t-distribution with a sparse precision matrix, which is shared across studies. This allows us to show that we can identify the co-expression matrix up to a scaling factor among other model parameters. Our method employs an Expectation-Maximization procedure for parameter estimation. Empirical evaluation on synthetic and gene expression data demonstrates our method's improved ability to learn the underlying graph structure compared to baseline methods.

1 Introduction

Over the past decades, advances in DNA sequencing technologies have led to significant advances in gene regulation research. These developments have provided deep insights into biological functions and disease processes. One notable example, which we will revisit later, is the comprehensive study of the bacterium *Bacillus subtilis*. This Gram-positive bacterium serves as a model organism for studying bacterial chromosome replication and cell differentiation. A substantial research endeavor has led to a continuous manual collection of biological findings about *Bacillus subtilis* regulation and gene functionality on the online platform *SubtiWiki* [24], providing a clearer and more precise understanding of its cellular processes. This underscores the importance of developing methods that facilitate this process by robustly identifying such gene-gene interactions in a vast collection of experimental data from multiple sources, such as different technologies and laboratories.

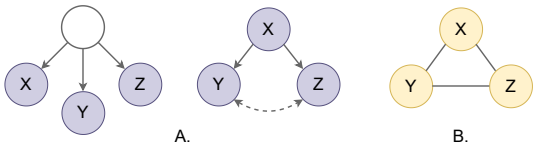

Figure 1: Two variations of the gene regulation of genes X, Y, Z (A) colored in purple and their corresponding co-expression network illustrated in (B) in yellow. In (A) (left), genes X, Y, Z are regulated by a common latent factor, such as another gene. The example in (A) (right) shows that gene X regulates both Y and Z. In addition, a bi-directional dashed line indicates potential confounding between genes Y and Z.

Biologically relevant gene-gene interactions are often represented by a gene co-expression network (GCN), which is an undirected graph where each node corresponds to a gene. Genes that are connected or positioned closely within the GCN belong to the same functional modules, indicating that

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

they work together to perform coordinated cellular activities. Therefore, constructing a GCN facilitates the understanding of gene regulation mechanisms. In this work, we aim to construct a GCN that closely resembles a gene regulatory network, considering only links that connect genes within the same regulatory network, such as regulator-regulated gene pairs or co-regulated genes (see Figure 1).

Inferring GCNs from data poses several challenges. First, the number of genes typically far exceeds the number of samples ($p \gg n$), making the problem high-dimensional. Second, experiments often provide only a small number of replicates per condition, sometimes as few as two, limiting the effectiveness of causal discovery algorithms in identifying gene regulatory relationships without external prior knowledge. Third, correlations exist not only between genes but also between samples, due to overlapping experimental designs and batch effects introduced by non-biological factors such as differences in technology or laboratory equipment. In response, current research in gene co-expression analysis often makes specific assumptions about the data generation model to deal with this complexity. This is typically represented by a noisy decomposition model: $\mathbf{X} = \mathbf{S}A + \mathbf{E}$, where $\mathbf{X} \in \mathbb{R}^{p \times n}$ is a gene expression matrix describing the activity of p genes across n different samples (experiments, patients, tissues, etc.), $\mathbf{S} \in \mathbb{R}^{p \times k}$ is the *gene loading matrix*, A is the *sample loading matrix*, and $\mathbf{E}$ is the additive noise. A common assumption is that GCNs can be reconstructed from the gene loadings, where gene clusters are identified from each latent vector, a column in the gene loadings matrix $\mathbf{S}$ (e.g. [20, 12, 27]).

This paper presents a novel probabilistic method for inferring complex network structures from high-dimensional data across multiple views. Unlike traditional approaches that rely primarily on clustering techniques or Gaussian models (e.g. [20, 12, 10, 15, 5, 11]), our method employs a matrix-variate t-distribution framework that extends TLASSO by Finegold and Drton [7]. We refer to this extended model as MVTLASSO, which captures the covariance at both the sample and variable levels in the multi-view setting. Key contributions of this work, besides the proposed model, include the formulation of identifiability guarantees for the model parameters, such as the sparse precision matrix, which we can identify up to a scalar multiple (see Section 2.1). For model estimation, we implement an Expectation-Maximization (EM) procedure, which is described in Section 3. We apply MVTLASSO to both synthetic datasets and real-world gene expression data to validate its effectiveness. Our empirical results in Section 4 show that MVTLASSO consistently demonstrates improved accuracy in reconstructing the underlying graph structures compared to baseline methods.

2 Robust Co-Expression Inference from non-i.i.d Samples

In this section, we introduce and justify our chosen generative model, which we will refer to as **MVTLASSO**, placing it within the broader context of known GCN inference methods. In Section 2.1, we present theoretical guarantees for recovering the true model parameters.

Our approach can be seen as an instance of ICA, where the latent components, or gene loadings, are divided into two categories: those used to construct the GCN, denoted by $\mathbf{S}$, and those considered noise, denoted by $\mathbf{Z}$, which do not contribute to the GCN inference. We infer the GCN from the sparse precision matrix Θ estimated from *all* "useful" gene loadings $\mathbf{S}$ across datasets (or *views*) that follow a multivariate t-distribution similar to [7]. More specifically, we make the following assumptions regarding the data generation process:

Definition 2.1. *Consider the scenario where we are given D different data sets $\mathbf{X}_d \in \mathbb{R}^{p \times n_d}$, which may come from different sources and follow the representation:*

$$\mathbf{X}_d = \mathbf{S}_d A_d + \mathbf{Z}_d B_d$$

where for each $d = 1, \ldots, D$ it holds

1. *$(A_d^\top | B_d^\top)^\top \in \mathbb{R}^{(k_d + r_d) \times n_d}$ have full row rank with $\operatorname{rank}(A_d) = k_d$ and $\operatorname{rank}(B_d) = n_d - k_d =: r_d$,*
2. *the columns of $\mathbf{S}_d \in \mathbb{R}^{p \times k_d}$ are mutually independent and follow a multivariate t-distribution, i.e. $\mathbf{S}_{d_{:,i}} \sim t_p(\nu, \mu_d, \Sigma)$ with $\nu > 2$ degrees of freedom and a sparse inverse dispersion matrix $\Theta := (\Sigma)^{-1}$ that has a prior distribution $p_\lambda(\Theta)$ with $\lambda > 0$ defined as*

$$p_\lambda(\Theta) \propto \exp\left(-\lambda \|\Theta\|_1\right) \quad \textit{with} \quad \|\Theta\|_1 = \sum_{i,j} |\Theta_{ij}|,$$

3. *the columns of* $\mathbf{Z}_d \in \mathbb{R}^{p \times r_d}$ *are noise random variables and are i.i.d multivariate t-distributed* $t_p(\nu, 0, \sigma_d^2 \mathbb{I}_p)$*, such that there is no* $\lambda \in \mathbb{R}$ *with* $\sigma_d^2 \mathbb{I}_p = \lambda \Sigma$,

4. *the latents* $\mathbf{S}_d$ *and noise matrix* $\mathbf{Z}_d$ *are independent.*

This perspective on Θ as a representation of the GCN closely aligns our work with that proposed by Stegle et al. [29] for the single view case. Compared to [29], we shift from a multivariate normal distribution to a multivariate t-distribution with sparse Θ. Although this moves away from the theoretical guarantees of conditional independence to a more relaxed condition of conditional uncorrelation, as outlined by Finegold and Drton [7], this approach provides more robust inference for unknown parameters, in this case, A_d, B_d, μ_d, Θ. This robustness is particularly beneficial in the presence of data contamination, a common challenge in the analysis of transcriptome data.

Finally, our model is reminiscent of dimensionality reduction methods, similar to the application of PCA, aiming to identify k_d components per dataset (or view) that capture the most significant signals from the data. The remaining components are considered as i.i.d. noise, following a multivariate t-distribution, which does not play a role in estimating the network structure, represented by Θ. A similar decomposition is proposed by Parsana et al. [23], where the authors show that removing the noise components after applying PCA improves the GCN inference of several algorithms.

2.1 Identifiability Guarantees

Next, we will present our theoretical guarantees for identifying model parameters from Definition 2.1, i.e $\{A_d, B_d, \mu_d, \sigma_d^2\}$, $d = 1, \ldots, D$, and Σ. We will show that the location μ_d and dispersion matrix Σ of the gene loadings, as well as the sample loadings A_d are identifiable up to the same constant across all views:

Proposition 2.1. *Let* $\mathbf{X}_1, \ldots, \mathbf{X}_D$ *with* $\mathbf{X}_d \in \mathbb{R}^{p \times n_d}$ *be random matrices with the following two representations:*

$$\mathbf{S}_d^{(1)} A_d^{(1)} + \mathbf{Z}_d^{(1)} B_d^{(1)} = \mathbf{X}_d = \mathbf{S}_d^{(2)} A_d^{(2)} + \mathbf{Z}_d^{(2)} B_d^{(2)},$$

where for $d = 1, \ldots, D$*, both representations* $A_d^{(1)} \in \mathbb{R}^{k_d^{(1)} \times n_d}, B_d^{(1)} \in \mathbb{R}^{(n_d - k_d^{(1)}) \times n_d}, \mathbf{S}_d^{(1)} \in \mathbb{R}^{p \times k_d^{(1)}}, \mathbf{Z}_d^{(1)} \in \mathbb{R}^{p \times (n_d - k_d^{(1)})}$ *and* $A_d^{(2)} \in \mathbb{R}^{k_d^{(2)} \times n_d}, B_d^{(2)} \in \mathbb{R}^{(n_d - k_d^{(2)}) \times n_d}, \mathbf{S}_d^{(2)} \in \mathbb{R}^{p \times k_d^{(2)}}, \mathbf{Z}_d^{(2)} \in \mathbb{R}^{p \times (n_d - k_d^{(2)})}$ *satisfy the properties of Definition 2.1. Then, for* $d = 1, \ldots, D$*,* $k_d^{(1)} = k_d^{(2)} =: k_d$*. Furthermore, there exist permutation matrices* $P_{A_1}, \ldots, P_{A_D}, P_{B_1}, \ldots, P_{B_D}$ *and constants* $c, c_1, \ldots, c_D > 0$ *such that:*

$$A_d^{(2)} = c P_{A_d} A_d^{(1)}, \qquad \Sigma_S^{(2)} = \frac{\Sigma_S^{(1)}}{c^2}, \qquad \mu_{S_d}^{(2)} = \frac{\mu_{S_d}^{(1)}}{c},$$

$$B_d^{(2)} = c_d P_{B_d} B_d^{(1)}, \qquad \Sigma_{Z_d}^{(2)} = \frac{\Sigma_{Z_d}^{(1)}}{c_d^2}, \qquad \mu_{Z_d}^{(2)} = \frac{\mu_{Z_d}^{(1)}}{c_d}.$$

In contrast to well-established results in the ICA literature [4, 14], which provide identifiability for the univariate case, we extend these results to multivariate elliptic distributions, as shown in Corollary B.1. Proposition 2.1 is a special case and a direct consequence of our more general results. Unlike in the single-view case, the multi-view setting allows us to achieve identifiability of the sample matrices A_d across views, up to a common scaling factor c.

3 Parameter Estimation

We begin by deriving the data likelihood, drawing inspiration from the ICA literature, e.g. the works of [13, 1]. Instead of making derivations with respect to A_d and B_d, we proceed in terms of the inverse of the concatenated matrix, denoted as $W_d = (A_d \mid B_d)^{-1}$. Consequently, the "unmixed" signal $\mathbf{Y}_d := \mathbf{X}_d W_d$ represents the estimates for the latent vectors $\mathbf{S}_{d_{:,i}}$ for $i = 1, \ldots, k_d$, and $\mathbf{Z}_{d_{:,i}}$ for $i = 1, \ldots, n_d - k_d$, up to some scaling and permutation as described in Proposition 2.1[1]. These

[1]Specifically, the first k_d columns of $\mathbf{Y}_d$ correspond to the estimates of $\mathbf{S}_d$, while the remaining $n - k_d$ columns correspond to the estimates of $\mathbf{Z}_d$.

signals follow a multivariate t-distribution. Thus, the likelihood for all views $\mathbf{X}_1, \ldots, \mathbf{X}_D$ is:

$$p(\mathbf{X}_1, \ldots, \mathbf{X}_D \mid \{ W_d, \mu_d, \sigma_d \}_{d=1}^D, \Sigma) = \prod_{d=1}^{D} p(\mathbf{X}_d) = \prod_{d=1}^{D} |\det W_d| \, p(\mathbf{X}_d \cdot W_d) \tag{1}$$
$$= \prod_{d=1}^{D} |\det W_d| \prod_{i=1}^{n_d} t_p \left(\mathbf{Y}_{d_{:,i}} \mid \nu, \rho_{d_i}, \Phi_{d_i} \right),$$

where $\Phi_{d_i} = \mathbb{1}_{\{i \leq k_d\}} \Sigma + \mathbb{1}_{\{i > k_d\}} \sigma_d^2 \mathbb{I}_p$ and $\rho_{d_i} = \mathbb{1}_{\{i \leq k_d\}} \mu_d$. Thus, the data likelihood is proportional to the product of the probabilities of $\sum_{d=1}^{D} n_d$ independent multivariate t-distributed vectors.

3.1 The Expectation-Maximization Procedure

Unfortunately, directly estimating the unknown parameters from (1) is infeasible. However, we can leverage the alternative representation of the multivariate t-distribution described in Theorem C.1, which is central to the EM procedure proposed by Liu and Rubin [18], Finegold and Drton [7]. For each random vector $\mathbf{Y}_{d_{:,i}}$, the corresponding generative process can equivalently be represented as:

$$\tau_{d_i} \sim \Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\right), \qquad \mathbf{Y}_{d_{:,i}} \sim \mathcal{N}(\rho_{d_i}, \Phi_{d_i}/\tau_{d_i}),$$

where the variables τ_{d_i} are unobserved. Thus, the complete data log-likelihood with unknown parameters $\boldsymbol{\gamma} := \{W_d, \mu_d, \sigma_d\}_{d=1}^D \cup \{\Sigma\}$ and random variables $\mathbf{X}_1, \ldots, \mathbf{X}_D$ and $\tau_1, \ldots \tau_D$ with $\tau_d := (\tau_{d_1}, \ldots, \tau_{d_{n_d}})$ is given by

$$l\left(\boldsymbol{\gamma};\ \{\mathbf{X}_d, \tau_d\}_{d=1}^D\right) \propto \sum_d \Bigg\{ \ln |\det W_d| + \sum_{i=1}^{n_d} \frac{1}{2} \ln \det \Phi_{d_i}^{-1} - \frac{\tau_{d_i}}{2} \operatorname{tr}\left(\Phi_{d_i}^{-1} \mathbf{Y}_{d_{:,i}} \mathbf{Y}_{d_{:,i}}^\top\right) \tag{2}$$
$$+ \tau_{d_i} \rho_{d_i}^\top \Phi_{d_i}^{-1} \mathbf{Y}_{d_{:,i}} - \frac{\tau_{d_i}}{2} \rho_{d_i}^\top \Phi_{d_i}^{-1} \rho_{d_i} \Bigg\},$$

where $\Phi_{d_i}^{-1} = \mathbb{1}_{\{i \leq k_d\}} \Theta + \mathbb{1}_{\{i > k_d\}} \frac{1}{\sigma_d^2} \mathbb{I}_p$ with $\Theta = (\Sigma)^{-1}$. The right side of (2) is linear in the latent variables τ_{d_i}. Thus, for the E-step it suffices to compute $\mathbb{E}[\tau_{d_i} \mid \mathbf{X}_d]$ for every $d = 1, \ldots, D$ and $i = 1, \ldots, n_d$. This can be derived directly by observing that the conditional distribution $p(\tau_{d_i} \mid \mathbf{X}_d) = p(\tau_{d_i} \mid \mathbf{Y}_{d_{:,i}})$ is given by

$$\tau_{d_i} \mid \mathbf{Y}_{d_{:,i}} \sim \Gamma\left(\frac{\nu + p}{2}, \frac{\nu + \delta(\mathbf{Y}_{d_{:,i}}, \rho_{d_i}, \Phi_{d_i})}{2}\right)$$

with

$$\delta(\mathbf{Y}_{d_{:,i}}, \rho_{d_i}, \Phi_{d_i}) = (\mathbf{Y}_{d_{:,i}} - \rho_{d_i})^\top \Phi_{d_i}^{-1} (\mathbf{Y}_{d_{:,i}} - \rho_{d_i}).$$

Consequently, for the conditional expectation we get: $\mathbb{E}[\tau_{d_i} \mid \mathbf{Y}_{d_{:,i}}] = \dfrac{\nu + p}{\nu + \delta(\mathbf{Y}_{d_{:,i}}, \rho_{d_i}, \Phi_{d_i})}$.

Hence, the EM procedure iterates through two main steps for each view d: 1) the estimation of τ_{d_i} while keeping ρ_{d_i}, Φ_{d_i}, and W_d fixed; and 2) the estimation of ρ_{d_i}, Φ_{d_i}, W_d, and $\Theta := (\Sigma)^{-1}$, where Θ is determined by solving the graphical lasso (GLASSO) problem as described by [8]. This method is designed to estimate sparse precision matrices in a multi-view setting. The EM procedure at step $t \geq 1$ is performed as follows:

E-step: For fixed estimated $\mu_d^{(t-1)}, \Sigma^{(t-1)}, \sigma_d^{(t-1)}$ and $W_d^{(t-1)}$ compute $\mathbb{E}[\tau_{d_i} \mid \mathbf{X}_d]$, i.e.

$$\mathbf{Y}_d^{(t-1)} = \mathbf{X}_d W_d^{(t-1)}, \qquad \tau_{d_i}^{(t)} = \frac{\nu + p}{\nu + \delta(\mathbf{Y}_{d_{:,i}}^{(t-1)}, \rho_{d_i}^{(t-1)}, \Phi_{d_i}^{(t-1)})}.$$

M-step: Solve the optimization problem:

$$\boldsymbol{\gamma}^{(t)} \in \arg\max_{\boldsymbol{\gamma}} l\left(\boldsymbol{\gamma};\ \{\mathbf{X}_d, \tau_d^{(t)}\}_{d=1}^D\right),$$

with $\boldsymbol{\gamma}^{(t)} = \{W_d^{(t)}, \mu_d^{(t)}, \sigma_d^{(t)}\}_{d=1}^D \cup \{\Sigma^{(t)}\}$ that leads to the following steps for all $d = 1, \ldots, D$:

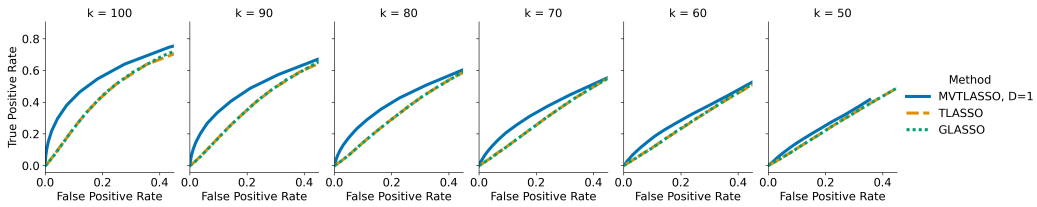

Figure 2: ROC curves summarizing the benchmark experiment on the data generated as described in Section 4.1 with a total of 100 sample loadings. Each curve represents the average result of 100 experiments. The number of signal loadings k varies in each experiment, as indicated in the subplot titles. The results show that MVTLASSO outperforms TLASSO and GLASSO.

1. Calculate $\mu_d^{(t)}, \Sigma^{(t)}$ and $\sigma_d^{(t)}$ for fixed $\tau_{d_i}^{(t)}$ and $\mathbf{Y}_d^{(t)}$

$$\mu_d^{(t)} = \frac{\sum_{i=1}^{k_d} \tau_{d_i}^{(t)} \mathbf{Y}_{d_{:,i}}^{(t-1)}}{\sum_{i=1}^{k_d} \tau_{d_i}^{(t)}}, \quad \Sigma^{(t)} = \frac{1}{\sum_d k_d} \sum_d \sum_{i=1}^{k_d} \tau_{d_i}^{(t)} \left(\mathbf{Y}_{d_{:,i}}^{(t-1)} - \mu_d^{(t)}\right) \left(\mathbf{Y}_{d_{:,i}}^{(t-1)} - \mu_d^{(t)}\right)^\top,$$

$$\sigma_d^{(t)} = \sqrt{\frac{1}{p(n-k_d)} \sum_{i=k_d+1}^{n} \sum_{l=1}^{p} \tau_{d_i}^{(t)} (\mathbf{Y}_{d_{l,i}}^{(t-1)})^2}$$

2. Estimate Θ via solving the GLASSO optimization problem for $\Sigma^{(t)}$ with penalty parameter $\lambda > 0$ given by:

$$\Theta^{(t)} \in \arg\min_{\Theta \succ 0} -\ln\det(\Theta) + \operatorname{tr}(\Sigma^{(t)}\Theta) + \lambda\|\Theta\|_1 \tag{3}$$

3. Estimate $W_d^{(t)}$ for fixed $\mu_d^{(t)}, \Sigma^{(t)}, \sigma_d^{(t)}$ and $\tau_{d_i}^{(t)}$:

$$W_d^{(t)} \in \arg\min_W \Big\{ \operatorname{tr}\left(\left(\mathbf{X}_d W - \boldsymbol{\mu}_d^{(t)}\right)^\top \Theta^{(t)} \left(\mathbf{X}_d W - \boldsymbol{\mu}_d^{(t)}\right) \mathcal{T}_1^{(t)}\right) + \frac{1}{(\sigma_d^{(t)})^2} \operatorname{tr}\left(W^\top \mathbf{X}_d^\top \mathbf{X}_d W \mathcal{T}_2^{(t)}\right) - \ln|\det W| \Big\}, \tag{4}$$

where $\boldsymbol{\mu}_d^{(t)} := (\underbrace{\mu_d^{(t)}, \ldots, \mu_d^{(t)}}_{k_d}, 0 \ldots, 0) \in \mathbb{R}^{p \times n_d}$, and $\mathcal{T}_1^{(t)}, \mathcal{T}_2^{(t)} \in \mathbb{R}^{n_d \times n_d}$ are diagonal matrices defined as $\mathcal{T}_1^{(t)} = \operatorname{diag}(\tau_{d_1}^{(t)}, \ldots, \tau_{d_{k_d}}^{(t)}, 0, \ldots, 0)$ and $\mathcal{T}_2^{(t)} = \operatorname{diag}(0, \ldots, 0, \tau_{d_{k_d+1}}^{(t)}, \ldots, \tau_{d_{n_d}}^{(t)})$

Details on the implementation of the EM procedure can be found in Appendix E.1. Code is available at `https://github.com/tpandeva/mvtlasso-code`.

4 Results

4.1 Simulated Data

We benchmark our method against **GLASSO** [8] and **TLASSO** [7] using simulated data whose generative model aligns with Definition 2.1 and follows a similar setup to that proposed by Finegold and Drton [7], with 200 variables and 100 samples. The sparse precision matrix Θ is generated as follows 1) off-diagonal entries Θ_{ij} with $i \neq j$ are sampled from $\{-1, 0, 1\}$ with probabilities $\{0.01, 0.98, 0.01\}$ 2) the diagonal entries are set to 1 plus the number of edges connected with the node, i.e. $\Theta_{ii} = 1 + \sum_j \mathbb{1}_{\{\Theta_{ij} \neq 0\}}$. Additionally, we set $\mu = 0$ and $\sigma = 1$ in all experiments. The sample loading matrices A and B have entries sampled according to standard normal distribution.

In the single view case ($D = 1$), we evaluated our model and several baselines on datasets with varying ratios of k signal loadings ($\mathbf{S}$) to r noise loadings ($\mathbf{Z}$): $100:0, 90:10, 80:20, 70:30, 60:40$, and $50:50$. We ran 100 experiments for each method for 50 sparsity parameters λ. For each experiment, we calculated the true positive and false positive rates individually and then averaged these for each sparsity parameter.

The aggregated results are shown in Figure 2. Across all scenarios indicated by the subplot titles, MVTLASSO, as expected, consistently outperforms the baselines. However, it is evident that the performance of all methods decreases as the proportion of noise loadings increases. Figure 3 illustrates that with an increasing number of views D, where the ratio of signal loadings k to noise loadings r is $50:50$, the performance of MVTLASSO improves significantly. Specifically, the aggregated true positive rate from 100 experiments increases relative to the false positive rate.

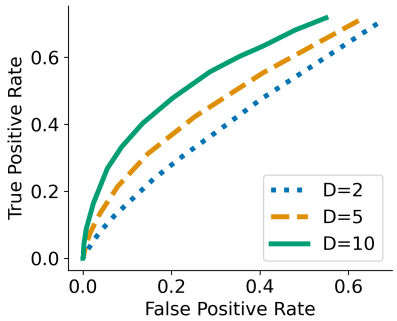

Figure 3: We set both $k = 50$ and $r = 50$ and varied $D = 2, 5, 10$. For each case, we ran 100 synthetic experiments while varying the sparsity parameter. The averaged ROC curves show that increasing the number of views (D) improves performance.

4.2 Gene Co-Expression Inference for Bacillus Subtilis

We revisit the motivational example of the GCN inference from *B. Subtilis* gene expression data. For this purpose, we use two well-controlled transcriptome data compendia. These datasets were collected using the *B. subtilis* strain BSB1, which contains 269 samples from 104 different experimental conditions [21], and the closely related strain PY79, which contains data from 38 unique experimental designs [3]. The *B. Subtilis* genome contains approximately 4100 genes, and for every transcriptome experiment, gene expressions from 3994 genes were obtained. Both datasets include a wide range of conditions, including growth in different media, competence, biofilm formation, swarming, different stress conditions, sporulation, and knockout experiments. The data were preprocessed as outlined in Appendix E.2. We further split each dataset into two approximately equal subsets of samples, ensuring they are as distinct as possible in their experimental design.

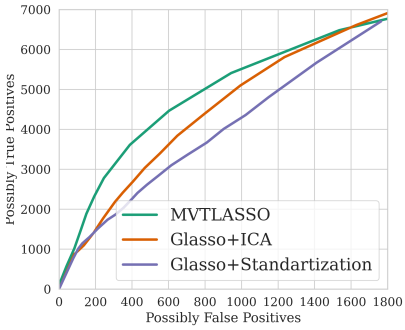

Figure 4: True positive vs. possibly false positive edges obtained via stability selection for various penalty parameters. The results demonstrate that MVTLASSO consistently infers more true positive edges across all settings.

We then used these four views to benchmark MVTLASSO against two other methods: GLASSO+ICA and GLASSO+Standardization, described in Appendix E.1. Unfortunately, the correct number of gene loadings k_d remains unknown. We estimated k_d following the approach in [23] prior to fitting the models. The fitting process for all methods includes stability selection as described by Meinshausen and Bühlmann [19] and detailed in Appendix E.1. In this approach, for each of 15 penalty parameters,"stable" edges are selected from 100 precision matrices, each estimated from bootstrapped samples containing 90% of all data. For each penalty parameter, we count the true positive edges, as verified against the ground truth data from *SubtiWiki*, as well as potential false positives from all selected edges. The true positive vs false positive counts for each method are shown in Figure 4. These results indicate that MVTLASSO consistently identifies more true positive edges across most penalty parameters compared to the other two methods.

5 Discussion

We introduced MVTLASSO, a robust method for inferring gene co-expression networks from high-dimensional gene expression data across multiple independent studies. Our approach effectively addresses the inherent complexity of gene expression data, including gene and sample correlations as well as batch effects, by modeling each dataset as a noisy linear mixture of gene loadings governed by a multivariate t-distribution with a sparse precision matrix. We employ an EM procedure for parameter estimation, supported by theoretical guarantees that ensure the identifiability of the model parameters. Empirical evaluations on both synthetic and real gene expression data have demonstrated the superior performance of MVTLASSO compared to baseline methods. Our method consistently shows improved accuracy in learning the underlying graph structures, underscoring its robustness and reliability.

A promising direction for future work is to develop a more efficient and reliable hyperparameter selection procedure. The selection of sample dimensions and noise loadings can be challenging and time-consuming due to the implemented EM procedure. In addition, incorporating available experimental metadata into the modeling process could provide further refinement and improve the overall performance of MVTLASSO.

References

[1] S.-i. Amari, A. Cichocki, and H. Yang. A new learning algorithm for blind signal separation. *Advances in Neural Information Processing Systems*, 8, 1995.

[2] R. B. Arellano-Valle and H. Bolfarine. On some characterizations of the t-distribution. *Statistics & Probability Letters*, 25(1):79–85, 1995.

[3] M. L. Arrieta-Ortiz, C. Hafemeister, A. R. Bate, T. Chu, A. Greenfield, B. Shuster, S. N. Barry, M. Gallitto, B. Liu, T. Kacmarczyk, et al. An experimentally supported model of the Bacillus subtilis global transcriptional regulatory network. *Molecular systems biology*, 11(11): 839, 2015.

[4] P. Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

[5] P. Danaher, P. Wang, and D. M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(2):373–397, 2014.

[6] K. W. Fang, S. Kotz, and K. W. Ng. *Symmetric Multivariate and Related Distributions*. Chapman and Hall/CRC, 2018.

[7] M. Finegold and M. Drton. Robust graphical modeling of gene networks using classical and alternative t-distributions. *The Annals of Applied Statistics*, 5(2A):1057–1080, 2011.

[8] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[9] J. Friedman, T. Hastie, and R. Tibshirani. *glasso: Graphical Lasso: Estimation of Gaussian Graphical Models*, 2019. URL `https://CRAN.R-project.org/package=glasso`. R package version 1.11.

[10] C. Gao, I. C. McDowell, S. Zhao, C. D. Brown, and B. E. Engelhardt. Context specific and differential gene co-expression networks via Bayesian biclustering. *PLoS computational biology*, 12(7):e1004791, 2016.

[11] J. Guo, E. Levina, G. Michailidis, and J. Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, 2011.

[12] S. Hochreiter, U. Bodenhofer, M. Heusel, A. Mayr, A. Mitterecker, A. Kasim, T. Khamiakova, S. Van Sanden, D. Lin, W. Talloen, et al. FABIA: factor analysis for bicluster acquisition. *Bioinformatics*, 26(12):1520–1527, 2010.

[13] A. Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, 10(3):626–634, 1999.

[14] A. M. Kagan, Y. V. Linnik, C. R. Rao, et al. *Characterization problems in mathematical statistics*. Wiley-Interscience, 1973.

[15] H. Kim and H. Park. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23(12):1495–1502, 2007.

[16] S. Kim, D. Kang, Z. Huo, Y. Park, and G. C. Tseng. Meta-analytic principal component analysis in integrative omics application. *Bioinformatics*, 34(8):1321–1328, 11 2017. ISSN 1367-4803.

[17] K.-A. Lê Cao, D. Rossouw, C. Robert-Granié, and P. Besse. A sparse PLS for variable selection when integrating omics data. *Statistical applications in genetics and molecular biology*, 7(1), 2008.

[18] C. Liu and D. B. Rubin. Ml estimation of the t distribution using em and its extensions, ecm and ecme. *Statistica Sinica*, pages 19–39, 1995.

[19] N. Meinshausen and P. Bühlmann. Stability Selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(4):417–473, 2010.

[20] G. E. Moran, V. Ročková, and E. I. George. Spike-and-slab Lasso biclustering. *The Annals of Applied Statistics*, 15(1):148 – 173, 2021.

[21] P. Nicolas, U. Mäder, E. Dervyn, T. Rochat, A. Leduc, N. Pigeonneau, E. Bidnenko, E. Marchadier, M. Hoebeke, S. Aymerich, et al. Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis. *Science*, 335(6072):1103–1106, 2012.

[22] T. Pandeva and P. Forré. Multi-View Independent Component Analysis for Omics Data Integration. In *2023 ICLR First Workshop on Machine Learning & Global Health*, 2023.

[23] P. Parsana, C. Ruberman, A. E. Jaffe, M. C. Schatz, A. Battle, and J. T. Leek. Addressing confounding artifacts in reconstruction of gene co-expression networks. *Genome biology*, 20 (1):1–6, 2019.

[24] T. Pedreira, C. Elfmann, and J. Stülke. The current state of SubtiWiki, the database for the model organism Bacillus subtilis. *Nucleic Acids Research*, 50(D1):D875–D882, 10 2021. ISSN 0305-1048.

[25] K. Rychel, A. V. Sastry, and B. O. Palsson. Machine learning uncovers independently regulated modules in the Bacillus subtilis transcriptome. *Nature Communications*, 11(1):1–10, 2020.

[26] W. Saelens, R. Cannoodt, and Y. Saeys. A comprehensive evaluation of module detection methods for gene expression data. *Nature Communications*, 9(1):1–12, 2018.

[27] A. V. Sastry, Y. Gao, R. Szubin, Y. Hefner, S. Xu, D. Kim, K. S. Choudhary, L. Yang, Z. A. King, and B. O. Palsson. The Escherichia coli transcriptome mostly consists of independently regulated modules. *Nature Communications*, 10(1):1–14, 2019.

[28] A. K. Smilde, I. Måge, T. Naes, T. Hankemeier, M. A. Lips, H. A. Kiers, E. Acar, and R. Bro. Common and distinct components in data fusion. *Journal of Chemometrics*, 31(7):e2900, 2017.

[29] O. Stegle, C. Lippert, J. M. Mooij, N. Lawrence, and K. Borgwardt. Efficient inference in matrix-variate Gaussian models with iid observation noise. *Advances in Neural Information Processing Systems*, 24, 2011.

A Related Work

Inferring GCNs from data can be very challenging, mainly due to hidden confounders and batch effects associated with the different data sources. In response, current research in gene co-expression analysis often makes specific assumptions about the data generation model to deal with this complexity. This is typically represented by a noisy decomposition model: $\mathbf{X} = \mathbf{S}A + \mathbf{E}$, where $\mathbf{X} \in \mathbb{R}^{p\times n}$ is a gene expression matrix describing the activity of p genes across n different samples (experiments, patients, tissues, etc.), $\mathbf{S} \in \mathbb{R}^{p\times k}$ is the *gene loading matrix*, A is the *sample loading matrix*, and $\mathbf{E}$ is the additive noise. These approaches can be broadly categorized into decomposition methods and their refinements, biclustering algorithms.

Decomposition methods, including Independent Component Analysis (ICA), Principal Component Analysis (PCA), and other variations of factor analysis, have shown remarkable effectiveness in identifying clusters of genes connected in the GCN. These methods are used to analyze single data sets [26, 25] as well as to integrate data from multiple studies [17, 28, 16, 22]. A common assumption is that GCNs can be reconstructed from the gene loadings, where gene clusters are identified from each latent vector, a column in the gene loadings matrix $\mathbf{S}$, usually by thresholding. Often, these clusters are assumed to represent sets of genes connected within the GCN and mapped to gene modules with a common function.

Biclustering algorithms aim to cluster genes and samples simultaneously by applying sparsity constraints to both gene and sample loadings, e.g., [20, 12, 10, 15], providing a principled approach for a two-fold clustering. This approach assumes that the sample loading matrix A will have a sparse pattern, i.e., only a small group of genes will deviate within a small subset of samples. These methods are particularly useful for subgroup analyses, such as classifying patients into different subtypes based on gene expression levels.

Despite their ability to cluster, all these methods do not model the relationships between clusters and thus do not provide a comprehensive strategy for inferring gene co-expression graphs. One exception is the Kronecker graphical LASSO approach by [29], which constructs a sparse graph structure while modeling sample covariance. However, this method has not been extended to handle multiple datasets collected from different labs and may lack robustness against data contamination. On the other hand, existing methods that use the graphical LASSO to infer GCNs from various data sources [5, 11] do not address the confounding variables in the experiments and assume that the data are independent and identically distributed.

B Identifiability

In our analysis, we will make use of the multivariate elliptical distributions, denoted by $E_p(\mu, \Sigma)$, whose density $f(x; \mu, \Sigma)$ is proportional to $f(x; \mu, \Sigma) \propto g((x-\mu)^\top \Sigma(x-\mu))$ for some measurable function g and a positive semi-definite dispersion matrix Σ and median μ. An example of such elliptical distributions is the Gaussian and multivariate t-distribution. First, we show that the sample loadings are identifiable up to scaling and permutation, provided that none of the gene loadings have Gaussian marginals. This result is an extension of Theorem 10.3 in [14] for the multivariate case:

Lemma B.1. *Let $\mathbf{X} \in \mathbb{R}^{p\times n}$ be a random matrix. Assume the following two representations of $\mathbf{X}$*

$$\mathbf{S}^{(1)}A^{(1)} = \mathbf{X} = \mathbf{S}^{(2)}A^{(2)},$$

with the following properties for $i = 1, 2$:

1. *$A^{(i)} \in \mathbb{R}^{k^{(i)}\times n}$ is a (non-random) matrix with a full row rank*
2. *$\mathbf{S}^{(i)} \in \mathbb{R}^{p\times k^{(i)}}$ is a random matrix such that the columns of $\mathbf{S}^{(i)}$ are mutually independent.*

If the i-th row of $A^{(1)}$ is not proportional to any row of $A^{(2)}$ then the i-th column of $\mathbf{S}^{(1)}$ has Gaussian distributed marginals. Additionally, if the i-th column of $\mathbf{S}^{(1)}$ follows an elliptical distribution, then it is multivariate Gaussian.

Proof of Lemma B.1

Proof. W.l.o.g. let $i = 1$. According to [14, Lemma 10.2.2] there exists a $n \times 2$ matrix H such that the matrices $C_1 = A^{(1)}H$ and $C_2 = A^{(2)}H$ of orders $k^{(1)} \times 2$ and $k^{(2)} \times 2$ respectively have the following property; the first row of C_1 is not proportional to any of the other rows of C_1 or to any of the rows of C_2.

Now consider the following algebraic relationship for $\mathbf{Y} = \mathbf{X}H$:

$$\mathbf{S}^{(1)}C_1 = \mathbf{Y} = \mathbf{S}^{(2)}C_2,$$

where $\mathbf{Y} \in \mathbb{R}^{p\times 2}$. For each row $r = 1, \dots p$ of $\mathbf{Y}$ we have the two equivalent representations

$$\mathbf{S}^{(1)}_{r,:}C_1 = \mathbf{Y}_{r,:} = \mathbf{S}^{(2)}_{r,:}C_2.$$

Thus, by [14, Lemma 10.2.4], it follows that $\mathbf{S}^{(1)}_{r,1}$ is Gaussian distributed because the first row of C_1 is not proportional to any of the other rows of C_1 or to the one of C_2. Consequently, this implies that the marginal distributions of $\mathbf{S}^{(1)}_{:,1}$ are Gaussians. Given that $\mathbf{S}^{(1)}_{:,1}$ is elliptical with the previous argument it follows that it is Gaussian [6]. □

Theorem B.1. *Let $\mathbf{X} \in \mathbb{R}^{p\times n}$ be a random matrix. Assume the following two representations of $\mathbf{X}$*

$$\mathbf{S}^{(1)}A^{(1)} = \mathbf{X} = \mathbf{S}^{(2)}A^{(2)}$$

with the following properties for $i = 1, 2$:

1. *$A^{(i)} \in \mathbb{R}^{k^{(i)}\times n}$ is a (non-random) matrix with full row rank, i.e. $\operatorname{rank}(A^{(i)}) = k^{(i)}$*
2. *$\mathbf{S}^{(i)} \in \mathbb{R}^{p\times k^{(i)}}$ is a random matrix such that*
   - *(a) The columns of $\mathbf{S}^{(i)}$ are mutually independent,*
   - *(b) For $k = 1, \dots, k^{(i)}$ the vectors $\mathbf{S}^{(i)}_{:,k}$ are distributed according to a non-Gaussian elliptical distribution $E_p(\mu^{(i)}, \Sigma^{(i)})$ with mean $\mu^{(i)}$ and a dispersion matrix $\Sigma^{(i)}$.*
   - *(c) Additionally, $\mathbf{S}^{(i)}_{:,k}$ the random vectors do not have Gaussian components.*

Then $k^{(1)} = k^{(2)} = k$ and there exist a permutation matrix $P = P(\rho) \in \mathbb{R}^{k\times k}$ given by $Pe_j = e_{\rho(j)}$ and a constant $c > 0$ such that:

$$A^{(2)} = cPA^{(1)}, \qquad \Sigma^{(2)} = \frac{\Sigma^{(1)}}{c^2} \qquad \mu^{(2)} = \frac{\mu^{(1)}}{c}.$$

Proof. Lemma B.1 establishes that each row of matrix $A^{(1)}$ is proportional to a row of $A^{(2)}$. Now if we assume that $k^{(1)} > k^{(2)}$ then there must be at least two distinct rows in $A^{(1)}$ that are proportional to the same row of $A^{(2)}$. This contradicts the assumption that both $A^{(1)}$ and $A^{(2)}$ have full row rank. Thus, it follows that $k^{(1)} = k^{(2)} =: k$ and there exist a permutation matrix $P \in \mathbb{R}^{k\times k}$ and an invertible diagonal matrix $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_k) \in \mathbb{R}^{k\times k}$ such that $A^{(2)} = \Lambda PA^{(1)}$.

Note that for the characteristic function of a matrix $\mathbf{S}$ that fulfills 2a) to c) for some mean μ and dispersion matrix Σ holds

$$\begin{aligned}
\chi_{\mathbf{S}}(\mathbf{t}) &= \mathbb{E}\left[\exp(i\operatorname{tr}(\mathbf{t}^\top\mathbf{S}))\right] = \mathbb{E}\left[\exp\left(i\sum_j \mathbf{t}_{:,j}^\top\mathbf{S}_{:,j})\right)\right] \\
&= \prod_j \chi_{\mathbf{S}_{:,j}}(\mathbf{t}_{:,j}) = \prod_j \chi_{\mathbf{S}_{:,1}}(\mathbf{t}_{:,j}) \\
&= \prod_j \exp\left(i\mathbf{t}_{:,j}^\top\mu\right)\psi\left(\mathbf{t}_{:,j}^\top\Sigma\mathbf{t}_{:,j}\right),
\end{aligned}$$

where ψ is the characteristic generator and $\mathbf{t} \in \mathbb{R}^{p\times k}$.

Let $\tilde{\mathbf{S}}^{(2)} = \mathbf{S}^{(2)}P$. Then we get for the characteristic functions of $\tilde{\mathbf{S}}^{(2)}$ and $\mathbf{S}^{(1)}$ for all $\mathbf{t} \in \mathbb{R}^{p\times k}$ $\chi_{\mathbf{S}^{(1)}}(\mathbf{t}) = \chi_{\tilde{\mathbf{S}}^{(2)}\Lambda}(\mathbf{t})$

$$\prod_j \exp\left(i\mathbf{t}_{:,j}^\top \mu^{(1)}\right)\psi_1\left(\mathbf{t}_{:,j}^\top \Sigma^{(1)}\mathbf{t}_{:,j}\right)$$
$$= \prod_j \exp\left(i\lambda_j \mathbf{t}_{:,j}^\top \mu^{(2)}\right)\psi_2\left(\lambda_j^2 \mathbf{t}_{:,j}^\top \Sigma^{(2)}\mathbf{t}_{:,j}\right),$$

where ψ_i is the characteristic generator corresponding to the $i-$the representation. Consequently, for each j with $\mathbf{t}_{:,j} = t \in \mathbb{R}^p$ and otherwise $\mathbf{t}_{:,r} = 0$ for all $r \neq j$ we get

$$\exp\left(it^\top \mu^{(1)}\right)\psi_1\left(t^\top \Sigma^{(1)}t\right)$$
$$= \exp\left(i\lambda_j t^\top \mu^{(2)}\right)\psi_2\left(\lambda_j^2 t^\top \Sigma^{(2)}t\right)$$

It follows that $\lambda_1 = \ldots = \lambda_k = c$ and $\mu^{(1)} = c\mu^{(2)}$, and $\Sigma^{(1)} = c^2\Sigma^{(2)}$. □

Next, we show that by imposing additional constraints on the gene loadings - in particular, requiring that they come from the same elliptic non-Gaussian multivariate distribution - it becomes possible to determine that the sample matrix, along with its locations and dispersion matrix, are identifiable up to a scalar:

Theorem B.2. *Let $\mathbf{X} \in \mathbb{R}^{p\times n}$ be a random matrix. Assume the following two representations of $\mathbf{X}$*

$$\mathbf{S}^{(1)}A^{(1)} + \mathbf{Z}^{(1)}B^{(1)} = \mathbf{X} = \mathbf{S}^{(2)}A^{(2)} + \mathbf{Z}^{(2)}B^{(2)}$$

with the following properties for $i = 1, 2$:

1. *$(A^{(i)\top}|B^{(i)\top})^\top \in \mathbb{R}^{(k^{(i)}+l^{(i)})\times n}$ is a (non-random) matrix with full row rank with $\mathrm{rank}(A^{(i)}) = k^{(i)}$ and $\mathrm{rank}(B^{(i)}) = l^{(i)} \leq n - k^{(i)}$*
2. *$\mathbf{S}^{(i)} \in \mathbb{R}^{p\times k^{(i)}}$ and $\mathbf{Z}^{(i)} \in \mathbb{R}^{p\times l^{(i)}}$ are random matrices such that for $i = 1, 2$ and $\mathbf{V}^{(i)} \in \{\mathbf{S}^{(i)}, \mathbf{Z}^{(i)}\}$*
   - *(a) The columns of $\mathbf{V}^{(i)}$ are mutually independent,*
   - *(b) The column vectors $\mathbf{V}_{:,k}^{(i)}$ are distributed according to a non-Gaussian elliptical distribution $E_p(\mu_V^{(i)}, \Sigma_V^{(i)})$ with location $\mu_V^{(i)}$ and a dispersion matrix $\Sigma_V^{(i)}$.*
   - *(c) Additionally, the random column vectors of $\mathbf{S}_d$ and $\mathbf{Z}_d$ do not have Gaussian components.*
3. *the latents $\mathbf{S}^{(i)}$ and noise matrix $\mathbf{Z}^{(i)}$ are independent and there exist no $\lambda \in \mathbb{R}$ such that $\mu_Z^{(i)} = \lambda\mu_S^{(i)}, \Sigma_Z^{(i)} = \lambda^2\Sigma_S^{(i)}$.*

Then $k^{(1)} = k^{(2)} = k$ and $l^{(1)} = l^{(2)} = l$ and exist permutation matrices $P_A, P_B \in \mathbb{R}^{k\times k}$ and constants $c_A, c_B > 0$ such that:

$$A^{(2)} = c_A P_A A^{(1)}, \qquad B^{(2)} = c_B P_B B^{(1)},$$
$$\Sigma_S^{(2)} = \frac{\Sigma_S^{(1)}}{c_A^2}, \qquad \mu_S^{(2)} = \frac{\mu_S^{(1)}}{c_A},$$
$$\Sigma_Z^{(2)} = \frac{\Sigma_Z^{(1)}}{c_B^2}, \qquad \mu_Z^{(2)} = \frac{\mu_Z^{(1)}}{c_B}.$$

Proof of Theorem B.2

Proof. According to Lemma B.1 each row of $(A^{(1)\top}|B^{(1)\top})^\top$ is proportional to a row of $(A^{(2)\top}|B^{(2)\top})^\top$. With similar arguments as above it holds that $k^{(1)} + l^{(1)} = k^{(2)} + l^{(2)}$.

Suppose that the $j-$th row of $A^{(1)}$ is proportional to the $r-$th row of $B^{(1)}$. It follows that that there exist a constant λ such that for all $t \in \mathbb{R}^p$:

$$\exp\left(it^\top \mu_S^{(1)}\right)\psi_1\left(t^\top \Sigma_S^{(1)} t\right) = \exp\left(i\lambda t^\top \mu_Z^{(2)}\right)\psi_2\left(\lambda^2 t^\top \Sigma_Z^{(2)} t\right),$$

i.e., $\mu_S^{(1)} = \lambda \mu_Z^{(2)}$ and $\Sigma_S^{(1)} = \lambda^2 \Sigma_Z^{(2)}$. Thus, $k^{(1)} = k^{(2)}$ and $l^{(1)} = l^{(2)}$. The rest follows from Theorem B.1. □

Corollary B.1. *Let $\mathbf{X}_1, \ldots, \mathbf{X}_D$ with $\mathbf{X}_d \in \mathbb{R}^{p \times n_d}$ be random matrices with the following two representations*

$$\mathbf{S}_d^{(1)} A_d^{(1)} + \mathbf{Z}_d^{(1)} B_d^{(1)} = \mathbf{X}_d = \mathbf{S}_d^{(2)} A_d^{(2)} + \mathbf{Z}_d^{(2)} B_d^{(2)}$$

with the following properties for $i = 1, 2$ and $d = 1, \ldots, D$:

1. *$(A_d^{(i)\top} | B_d^{(i)\top})^\top \in \mathbb{R}^{(k_d^{(i)} + l_d^{(i)}) \times n_d}$ is a (non-random) matrix with full row rank:*

$$\operatorname{rank}(A_d^{(i)}) = k_d^{(i)}, \qquad \operatorname{rank}(B_d^{(i)}) = l_d^{(i)} \leq n_d - k_d^{(i)},$$

2. *the columns of $\mathbf{S}_d^{(i)}$ are independent and are distributed according to a non-Gaussian elliptical distribution $E_p(\mu_{S_d}^{(i)}, \Sigma_{S_d}^{(i)})$ with location $\mu_{S_d}^{(i)}$ and a dispersion matrix $\Sigma_S^{(i)} := \Sigma_{S_1}^{(i)} = \ldots = \Sigma_{S_D}^{(i)}$.*

3. *the columns of $\mathbf{Z}_d^{(i)}$ are noise random variables and are i.i.d non-Gaussian elliptical distributed $E_p(\mu_{Z_d}^{(i)}, \Sigma_{Z_d}^{(i)})$ with location $\mu_{Z_d}^{(i)}$ and a dispersion matrix $\Sigma_{Z_d}^{(i)}$. Furthermore, for each d there exist no $\lambda \in \mathbb{R}$ such that $\mu_{Z_d}^{(i)} = \lambda \mu_S^{(i)}, \Sigma_{Z_d}^{(i)} = \lambda^2 \Sigma_S^{(i)}$.*

4. *the latents $\mathbf{S}_d^{(i)}$ and noise matrix $\mathbf{Z}_d^{(i)}$ are mutually independent.*

5. *Additionally, the random column vectors of $\mathbf{S}_d$ and $\mathbf{Z}_d$ do not have Gaussian components.*

Then, for $d = 1, \ldots, D$, $k_d^{(1)} = k_d^{(2)} = k_d$ and $l_d^{(1)} = l_d^{(2)} = l_d$. Furthermore, there exist permutation matrices $P_{A_1}, \ldots, P_{A_D}, P_{B_1}, \ldots, P_{B_D}$ and constants $c_A, c_{B_1}, \ldots, c_{B_D} > 0$ such that:

$$A_d^{(2)} = c_A P_{A_d} A_d^{(1)}, \qquad B_d^{(2)} = c_{B_d} P_{B_d} B_d^{(1)},$$
$$\Sigma_S^{(2)} = \frac{\Sigma_S^{(1)}}{c_A^2}, \qquad \mu_{S_d}^{(2)} = \frac{\mu_{S_d}^{(1)}}{c_A},$$
$$\Sigma_{Z_d}^{(2)} = \frac{\Sigma_{Z_d}^{(1)}}{c_{B_d}^2}, \qquad \mu_{Z_d}^{(2)} = \frac{\mu_{Z_d}^{(1)}}{c_{B_d}}.$$

Proof of Corollary B.1 Theorem B.2 guarantees the identifiability results for each view separately, i.e. for each $d = 1, \ldots, D$ there exist permutation matrices P_{A_d}, P_{B_d} and constants $c_{A_d}, c_{B_d} > 0$ such that:

$$A_d^{(2)} = c_{A_d} P_{A_d} A_d^{(1)}, \qquad B_d^{(2)} = c_{B_d} P_{B_d} B_d^{(1)},$$
$$\Sigma_S^{(2)} = \frac{\Sigma_S^{(1)}}{c_{A_d}^2}, \qquad \mu_{S_d}^{(2)} = \frac{\mu_{S_d}^{(1)}}{c_{A_d}},$$
$$\Sigma_{Z_d}^{(2)} = \frac{\Sigma_{Z_d}^{(1)}}{c_{B_d}^2}, \qquad \mu_{Z_d}^{(2)} = \frac{\mu_{Z_d}^{(1)}}{c_{B_d}}.$$

It follows that for all $d = 1, \ldots, D$:

$$\Sigma_S^{(2)} = \frac{\Sigma_S^{(1)}}{c_{A_d}^2}.$$

Thus, $c_A := c_{A_1} = \ldots = c_{A_D}$.

C Dependence Structure and Properties of the Multivariate t-Distribution

C.1 Alternative Generative Model for the Multivariate t-Distribution

The probability density function of the multivariate t-distribution with ν degrees of freedom, mean vector μ, and scale matrix $\mathbf{\Sigma}$ in p dimensions is given by:

$$f(\mathbf{x}) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}}\left(1+\frac{1}{\nu}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right)^{-\frac{\nu+p}{2}}$$

where:

- $\mathbf{x}$ is the variable vector,
- μ is the mean vector,
- $\mathbf{\Sigma}$ is the scale matrix,
- ν is the degrees of freedom,
- Γ is the gamma function.

The following result is central to the EM procedure and is shows that the multivariate t-distribution can be expressed by means of the multivariate normal distributed random variable and Gamma distributed random variable:

Theorem C.1 ([2]). *Let $S \sim t_p(\nu, \mu, \Sigma)$ for some mean μ and positive semi-definite matrix Σ. Then, there exist random variables τ and N that follow Gamma distribution $\Gamma(\frac{\nu}{2}, \frac{\nu}{2})$ and a Gaussian distribution $\mathcal{N}(0, \Sigma)$, respectively, such that $S \sim \mu + N/\sqrt{\tau}$.*

C.2 Dependence Relationship between the Genes in the GCN

As previously discussed, we use the precision matrix Θ to construct the GCN. Specifically, consider a graph $G = (V, E)$, where $V = \{1, \ldots, p\}$ represents the set of observed genes and E is a collection of edges between pair of nodes (or genes) i and j for which the corresponding entry Θ_{ij} is non-zero.

An interesting aspect is understanding the types of (in)dependencies encoded by this graph structure. For context, in Gaussian models, the absence of an edge between two nodes i and j implies conditional independence between them, given the remaining nodes. However, this direct implication does not translate to multivariate t-distributions. Instead, a weaker concept of dependence, conditional uncorrelation, applies, as discussed in [7] for the single view case:

Theorem C.2 ([7]). *Let $S \sim t_p(\nu, \mu, \Sigma)$. Σ is a positive definite matrix with $(\Sigma^{-1})_{ij} = 0$ for indices $i \neq j$ corresponding to the non-edges in the graph G. If two nodes i and j are separated by a set of nodes C in G, then S_i and S_j are conditionally uncorrelated given S_C.*

While Theorem C.2 shows that conditional uncorrelation can be derived from the graph structure, it leaves open the question of whether multivariate t-distributions can be factorized according to any Bayesian network. The following result addresses this issue by showing that the only Bayesian network compatible with the multivariate t-distribution is a fully connected DAG:

Lemma C.1. *Let $G = (V, E)$ be a DAG with vertices $V = \{1, \ldots, p\}$. Furthermore, the joint distribution of the corresponding variables $S_1, \ldots, S_p$ is multivariate t-distribution $t_p(\nu, \mu, \Sigma)$ with $0 < \nu < \infty$. Let $\mathrm{pa}(k) \subseteq V \setminus \{k\}$ denote the set of parents of node k. Then, the following holds $P(S_1, \ldots, S_p) = \prod_{k=1}^{p} P(S_k \mid S_{\mathrm{pa}(k)})$ iff there exists an ordering $S_{\tau(1)}, \ldots, S_{\tau(p)}$ such that $\mathrm{pa}(\tau(k)) = \{\tau(1), \ldots, \tau(k-1)\}$, i.e. the graph is fully connected.*

Remark C.1. *(a) Lemma C.1 suggests that from the estimated Θ, we can infer only conditional uncorrelation between the genes, not conditional independence. However, this result does not contradict the GCN definition used in this work, as detailed in Section 1, which is based on correlation rather than statistical independence.*

(b) According to Theorem C.2, the reconstructed GCN should exclude edges between genes that are conditionally uncorrelated given rest of the genes. This implies that co-regulated genes

will not be connected in the GCN, as they become conditionally uncorrelated when conditioned on their regulators. However, this holds only in the absence of confounding factors—whether observed or unobserved—such as external stimuli that might influence gene regulation as part of the experimental design. Our current approach does not account for these elements of the experimental design, which could potentially refine the GCN. We leave this consideration for future work.

Proof of Lemma C.1

Proof. " $\Leftarrow$ " This direction follows directly from the chain rule of probabilities.

" $\Rightarrow$ " Assume that the DAG is not fully connected, i.e. there exist sets $A, B, C \subset V,\ A \neq \emptyset, B \neq \emptyset$ such that the random variables S_A and S_B are d-separated given S_C ($S_A \perp_G S_B | S_C$). Thus, it follows that $S_A \perp\!\!\!\perp S_B \mid S_C$ which implies that $p(S_A \mid S_B, S_C) = p(S_A | S_C)$.

According to [2] the joint distribution of S_A, S_B, S_C, their conditionals and marginals follow a multivariate t-distribution. More precisely, let $d = |A| + |B| + |C|,\ \mu_d = (\mu_A^\top, \mu_B^\top, \mu_C^\top)^\top,\ \Sigma = \Sigma_{(A,B,C),(A,B,C)}$, then $S_d = (S_A, S_B, S_C) \sim t_d(\nu, \mu_d, \Sigma)$. Furthermore, for the conditional distributions we have

$$S_A \mid S_B, S_C \sim t_{|A|}\left(\nu + |B| + |C|, \mu_{A|B,C}, \frac{\nu + d_{B,C}}{\nu + |B| + |C|}\Sigma_{A|B,C}\right) \tag{5}$$

$$S_A \mid S_C \sim t_{|A|}\left(\nu + |C|, \mu_{A|C}, \frac{\nu + d_C}{\nu + |C|}\Sigma_{A|C}\right)$$

Then, it follows that $\nu + |B| + |C| = \nu + |B| + |C|$ which implies that $|B| = 0$. □

D Parameter Inference: Background

D.1 Graphical LASSO

The Graphical lasso (GLASSO) is a maximum likelihood estimator for inferring graph structure within high-dimensional multivariate normal distributed data through estimating a sparse precision matrix [8]. More precisely, let $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ be a collection of n i.i.d. samples distributed according to the multivariate normal distribution $\mathcal{N}(0, \Theta^{-1})$, where $\Theta^{-1} \in \mathbb{R}^{p \times p}$ is the covariance matrix and its inverse Θ known as the precision matrix. The underlying undirected graph structure among the variables can be inferred directly from the precision matrix: a non-zero entry Θ_{ij} indicates an undirected edge between the i-th and j-the variables in the multivariate vector. GLASSO estimates Θ by maximizing the posterior distribution of $\mathbf{X}$ given $\Theta := \Sigma^{-1}$ which is proportional to

$$p(\mathbf{X}, \Theta) = p_\lambda(\Theta) \prod_{i=1}^{n} \mathcal{N}(\mathbf{X}_i | \mu, \Theta^{-1}) \quad \text{where} \quad \Theta \succ 0.$$

The prior $p_\lambda(\Theta)$ on the positive-definite matrices Θ parametrized by $\lambda > 0$ is defined as

$$p_\lambda(\Theta) \propto \exp\left(-\lambda \|\Theta\|_1\right) \quad \text{with} \quad \|\Theta\|_1 = \sum_{i,j} |\Theta_{ij}|.$$

Thus, the MLE problem that GLASSO solves can be formalized as follows

$$\max_{\Theta \succ 0} \ln p(\mathbf{X}, \Theta) \equiv \min_{\Theta \succ 0} -\ln\det(\Theta) + \operatorname{tr}(\hat{\Sigma}\Theta) + \lambda \|\Theta\|_1, \tag{6}$$

where S is the empirical covariance matrix, $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^\top$, and $\bar{\mathbf{X}}$ is the empirical mean. Intuitively, the parameter λ controls the sparsity level of the precision matrix Θ. Specifically, selecting a higher value for λ leads to sparser precision matrix estimates.

D.2 Student's t-Lasso

The accuracy of graph inference can be significantly compromised by deviations from the normal distribution assumption. To address this robustness issue, [7] propose an alternative to GLASSO for

inferring graph structure of multivariate Student's t-distribution which we call TGLASSO. Consider the setting from above, where we are given a collection of n i.i.d samples $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$. Then the joint distribution of the data $\mathbf{X}$ and precision matrix Θ is given by

$$p(\mathbf{X}, \Theta) = p_\lambda(\Theta) \prod_{i=1}^{n} t_{\nu,p}(\mathbf{X}_i | \mu, \Theta^{-1}) \quad \text{where} \quad \Theta \succ 0,$$

where the density function of the Student's t-distribution $t_{\nu,p}(\mu, \Theta^{-1})$ is given by

$$\frac{\Gamma\left((\nu+p)/2\right) \det \Theta^{1/2}}{(\pi\nu)^{p/2}\Gamma(\nu/2)\left(1+\delta\left(\mathbf{x};\mu,\Theta\right)/\nu\right)^{(\nu+p)/2}}$$

with

$$\delta\left(\mathbf{x};\mu,\Theta\right) = (\mathbf{x}-\mu)^\top \Theta(\mathbf{x}-\mu), \quad \mathbf{x} \in \mathbb{R}^p.$$

Estimating the precision matrix in this setting is not tractable, and [7] propose an Expectation-Maximization procedure for estimating Θ by exploiting the following generative model with latent variables $\mathbf{Z}_i$ and τ_i for each sample $\mathbf{X}_i$

$$\begin{aligned} \mathbf{Z}_i &\sim \mathcal{N}(0, \Theta^{-1}) \\ \tau_i &\sim \Gamma(\nu/2, \nu/2) \\ \mathbf{X}_i := \mu + \mathbf{Z}_i/\sqrt{\tau}_i &\sim t_{\nu,p}(\mu, \Theta^{-1}). \end{aligned}$$

The proposed EM procedure operates under the assumption that τ_i's are latent variables and that $\mathbf{X}_i|\tau_i \sim \mathcal{N}(\mu, (\tau_i\Theta)^{-1})$. This process iterates through two main steps: 1) Estimating the τ_i for fixed μ and Θ^{-1} and 2) Estimating μ and Θ^{-1}, where Θ is a solution to the GLASSO problem in Equation (6) for an empirical covariance matrix of the estimated $\mathbf{Z}$. More precisely, at step $t \geq 0$ the EM procedure becomes

E-step: For fixed estimated $\mu^{(t-1)}$ and $\Theta^{(t-1)}$ compute

$$\tau_i^{(t)} = \frac{\nu + p}{\nu + \delta\left(\mathbf{X}_i; \mu^{(t-1)}, \Theta^{(t-1)}\right)}$$

M-step: Calculate $\mu^{(t)}$ and $\Sigma^{(t)}$

$$\mu^{(t)} = \frac{\sum_{i=1}^n \tau_i^{(t)} \mathbf{X}_i}{\sum_{i=1}^n \tau_i^{(t)}} \qquad \Sigma^{(t)} = \frac{1}{n}\sum_{i=1}^n \tau_i^{(t)} \left(\mathbf{X}_i - \mu^{(t)}\right)\left(\mathbf{X}_i - \mu^{(t)}\right)^\top \tag{7}$$

Estimate $\Theta^{(t)}$ via solving the GLASSO optimization problem

$$\Theta^{(t)} \in \operatorname*{arg\,min}_{\Theta \succ 0} - \ln\det(\Theta) + \operatorname{tr}(\Sigma^{(t)}\Theta) + \lambda\|\Theta\|_1$$

E Experiments

E.1 Implementation

Implementation of MVTLASSO

1. The optimization problem in step 3 is convex if $\{W_d\}_{d=1}^D$ are positive semidefinite matrices. In the general case, however, we only require $\{W_d\}_{d=1}^D$ to be invertible, which increases the complexity of solving equation 4 . By treating the datasets $\mathbf{X}_1, \ldots, \mathbf{X}_D$ as instances of Independent Component Analysis (ICA), we can reformulate the original optimization problem equation 4 . Using FastICA [13] or a similar ICA algorithm, we obtain transformed datasets $\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_D$, which serve as estimates of $\mathbf{Y}_1, \ldots, \mathbf{Y}_D$, up to permutation matrices $P_1, \ldots, P_D$ and scaling factors. Consequently, replacing the raw data with $\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_D$ transforms the original optimization problem equation 4 into estimating each permutation matrix P_d. In our implementation, we optimize for P_d with an additional relaxation, allowing P_d to be orthogonal, which accelerates the computation.

2. We initialize the parameter W_0 as detailed in the previous remark (a), and the parameters μ_d, σ_d and Θ using a few iterations (not necessarily to convergence) of TLASSO by Finegold and Drton [7].
3. The steps in the M-step are interdependent, i.e., steps 1 and 2 are based on the inverse sample matrix W_d. Therefore, it is possible to iterate steps 1 to 3 multiple times. In our implementation, however, we perform only a single iteration.
4. Our model relies on several hyperparameters, including the number of multivariate t-distributed vectors k_d used for graph inference and the penalty parameters λ and γ. Ideally, these parameters could be determined by cross-validation. However, our EM procedure involves a GLASSO step in each iteration, which is computationally intensive. Therefore, we preselect the number of components prior to the parameter estimation process as described by Parsana et al. [23].

Implementation of baselines For GLASSO, we used the implementation available in the R package [9]. The variants GLASSO+Standardization and GLASSO+ICA include preliminary steps where the samples are subjected to standardization and ICA, respectively, before GLASSO is applied for precision matrix estimation. It is important to note that neither baseline includes a dimensionality reduction step.

Stability Selection The fitting procedure for all GLASSO-based methods makes use of stability selection by Meinshausen and Bühlmann [19] with a predefined range of penalty parameters. The steps of the procedure are outlined as follows:

1. The data is repeatedly subsampled by selecting 90% of all samples per view $N = 100$ times. For each subsample, the selected GCN inference method is applied using the predefined set of penalty parameters, Λ.
2. The outcomes for each penalty parameter are gathered in the selection probability matrix Π_λ, where $(\Pi_\lambda)_{ij}$ represents the proportion of the N precision matrices $\hat{\Theta}^{(1)}, \ldots, \hat{\Theta}^{(N)}$ indicating a nonzero edge between nodes i and j, i.e. $(\Pi_\lambda)_{ij} = \frac{\sum_l \mathbb{1}_{\{\hat{\Theta}_{ij}^{(l)} \neq 0\}}}{N}$.
3. We select the edges whose selection probability exceeds 50% for each penalty parameter.
4. The final graph can be constructed by collecting all edges inferred from the range of penalty parameters. This step is not conducted in our analysis.

The primary benefit of stability selection, as outlined by [19], is that it can reduce the risk of false positives, i.e., incorrectly identifying edges in the network. By requiring that an edge be consistently identified across many subsamples of the data, stability selection ensures that the edges selected are robust and not the result of random variations in the data.

E.2 Data Preprocessing

The dataset BSB1 is preprocessed following the method suggested by Rychel et al. [25]. Specifically, three samples (S3_3, G+S_1, and Mt0_2) were removed to ensure that the Pearson correlation between biological replicates was at least 0.9. Furthermore, we centered the data by subtracting the mean gene values in the M9 exponential growth condition. We used the preprocessed PY79 dataset by Arrieta-Ortiz et al. [3]. BSB1 and PY79 samples are then centered and rescaled before applying any graph inference procedures. We selected genes that are present in both datasets. In addition, we have split both datasets into two subsets of samples with experimental designs that are as different as possible to simulate four views instead of two. A link to the datasets will be provided after submission.