Inferring independent sets of Gaussian variables after thresholding correlations

Arkajyoti Saha\textsuperscript{1}\textsuperscript{*}, Daniela Witten\textsuperscript{1,2}\textsuperscript{†}, Jacob Bien\textsuperscript{3}\textsuperscript{‡}

\textsuperscript{1} Department of Statistics, University of Washington
\textsuperscript{2} Department of Biostatistics, University of Washington
\textsuperscript{3} Department of Data Sciences and Operations, University of Southern California

Abstract

We consider testing whether a set of Gaussian variables, selected from the data, is independent of the remaining variables. We assume that this set is selected via a very simple approach that is commonly used across scientific disciplines: we select a set of variables for which the correlation with all variables outside the set falls below some threshold. Unlike other settings in selective inference, failure to account for the selection step leads, in this setting, to excessively conservative (as opposed to anti-conservative) results. Our proposed test properly accounts for the fact that the set of variables is selected from the data, and thus is not overly conservative. To develop our test, we condition on the event that the selection resulted in the set of variables in question. To achieve computational tractability, we develop a new characterization of the conditioning event in terms of the canonical correlation between the groups of random variables. In simulation studies and in the analysis of gene co-expression networks, we show that our approach has much higher power than a “naive” approach that ignores the effect of selection.

Keywords: Selective inference, gene co-expression network, canonical correlation analysis, correlation thresholding.

\textsuperscript{*}Corresponding email: asaha3@uw.edu, AS, DW, and JB were supported by NIH Grant R01GM123993
\textsuperscript{†}DW was supported by Simons Investigator Award No. 560585, NIH Grant R01EB026908, and NIH Grant R01DA047869
\textsuperscript{‡}JB was supported by NSF CAREER Award DMS-1653017
1 Introduction

A vast statistical literature focuses on estimation of, and inference on, the sparsity structure of a covariance or correlation matrix (see, e.g. Bickel and Levina (2008); Drton and Perlman (2008); Lam and Fan (2009); Rothman et al. (2009); Bien and Tibshirani (2011); Cai and Yuan (2012); Liu et al. (2012); Rothman (2012); Xue et al. (2012); Zhao et al. (2012); Liu et al. (2014); Serra et al. (2018) and references therein). In this article, we focus on a particular type of sparsity in which the matrix has a block diagonal structure (up to a permutation of the rows and columns). This type of sparsity arises in a number of scientific fields. For example, in the study of gene co-expression networks, discovery of independent “gene sets” plays a pivotal role (e.g. Freeman et al. 2007). In genetics, discovery of haplotype blocks often reduces to partitioning the linkage disequilibrium matrix into approximately independent blocks (Patil et al., 2001). In neuroscience, it is of interest to discover correlated brain regions in brain functional connectivity networks (Bordier et al., 2017).

We consider a simple approach to estimating block diagonal structure in a correlation matrix. Our paper’s primary contribution is a novel approach to testing whether two sets of variables that appear in different estimated blocks are, in fact, uncorrelated. For example, is an estimated gene set in a gene co-expression network really uncorrelated with the remaining genes? This is a challenging question because the same data is used both to estimate the gene set and to test a hypothesis based on this estimated gene set. The problem falls into the selective inference framework, which has been studied in various other contexts by Fithian et al. (2014); Lee and Taylor (2014); Loftus and Taylor (2015); Lee et al. (2016); Yang et al. (2016); Tibshirani et al. (2016); Suzumura et al. (2017); Hyun et al. (2018); Tian et al. (2018); Hyun et al. (2018); Jewell et al. (2022); Chen and Bien (2020); Gao et al. (2022); Neufeld et al. (2021); Chen and Witten (2022), among others. However, unlike those other papers, in our setting, failure to account for double-use of data leads to a loss of power rather than loss of type I error control.

The rest of this paper is organized as follows. In Section 2, we briefly review some results related to testing hypotheses associated with submatrices of a covariance matrix, define the notion of selective type I error in this context, and then provide a general expression for constructing p-values that are valid in this selective sense. In Section 3, we present a data-adaptive procedure for estimating diagonal blocks in a correlation matrix, and consider testing the null hypothesis of no correlation between an estimated block and the remaining variables. Furthermore, we show that the event corresponding to the selection of a particular hypothesis can be characterized analytically. In Section 4, we apply this characterization to develop a selective inference approach to test the null hypothesis of interest. Simulation results in Section 5 and an analysis of gene co-expression networks in Section 6 demonstrate the value of the proposed approach. We close with a discussion in Section 7.
2 Problem definition

2.1 Hypothesis testing for pre-specified groups of variables

For a positive definite \( p \times p \) matrix \( \Sigma \), consider a data matrix consisting of \( n > p \) independent observations of a \( p \)-dimensional multivariate Gaussian,

\[
X \sim \mathcal{MN}_{n \times p}(0, I_n, \Sigma_{p \times p}).
\]

(1)

For a predetermined subset \( \mathcal{P} \subset \{1, 2, \ldots, p\} \) of variables, consider testing for dependence between the variables in \( \mathcal{P} \) and the variables in its complement \( \mathcal{P}^c := \{1, 2, \ldots, p\} \setminus \mathcal{P} \), i.e.

\[
H_0^\mathcal{P} : \Sigma_{\mathcal{P}, \mathcal{P}^c} = 0 \text{ versus } H_1^\mathcal{P} : \Sigma_{\mathcal{P}, \mathcal{P}^c} \neq 0,
\]

(2)

where \( \Sigma_{\mathcal{P}_1, \mathcal{P}_2} \) denotes the submatrix of \( \Sigma \) with rows in \( \mathcal{P}_1 \) and columns in \( \mathcal{P}_2 \). Given a realization \( x \in \mathbb{R}^{n \times p} \) of \( X \), the classical likelihood ratio test (LRT) of (2) yields the p-value

\[
plrt(x; \mathcal{P}) := P_{H_0^\mathcal{P}} \left( \frac{|S(X)|}{|S_{\mathcal{P}, \mathcal{P}}(x)||S_{\mathcal{P}^c, \mathcal{P}^c}(x)|} \leq \frac{|S(x)|}{|S_{\mathcal{P}, \mathcal{P}}(x)||S_{\mathcal{P}^c, \mathcal{P}^c}(x)|} \right),
\]

(3)

where \( S(x) = \frac{1}{n} (x - 1_n x^\top) \mathcal{P} (x - 1_n x^\top) \) is the sample covariance matrix of \( x \), \( \bar{x} \in \mathbb{R}^p \) denotes the sample mean of \( x \), and \( |M| \) denotes the determinant of the matrix \( M \) (see (3) of Section 9.8 in Anderson 2003). Here, \( S_{\mathcal{P}_1, \mathcal{P}_2} \) represents the submatrix of \( S \) with rows in \( \mathcal{P}_1 \) and columns in \( \mathcal{P}_2 \). The compact singular value decomposition (SVD) of the cross-covariance of decorrelated (“whitened”) versions of \( x_\mathcal{P} \) and \( x_{\mathcal{P}^c} \) (defined as the submatrices of \( x \) with columns in \( \mathcal{P} \) and \( \mathcal{P}^c \), respectively) can be written as

\[
S_{\mathcal{P}, \mathcal{P}^c}^W(x) := (S_{\mathcal{P}, \mathcal{P}}(x))^{-\frac{1}{2}} S_{\mathcal{P}, \mathcal{P}^c}(x) (S_{\mathcal{P}^c, \mathcal{P}^c}(x))^{-\frac{1}{2}} \text{ SVD } \hat{A}^\mathcal{P}(x) \hat{\Lambda}^\mathcal{P}(x) \left( \hat{\Gamma}^\mathcal{P}(x) \right)^\top,
\]

(4)

where \( \hat{\Lambda}^\mathcal{P}(x) := \text{diag} \left( \hat{\lambda}_1(x), \hat{\lambda}_2(x), \ldots, \hat{\lambda}_{r(\mathcal{P})}(x) \right) \) is a square diagonal matrix of \( r(\mathcal{P}) := \min \{ \text{card} (\mathcal{P}), \text{card} (\mathcal{P}^c) \} \) non-negative singular values in non-increasing order \( \hat{\lambda}_1(x) \geq \hat{\lambda}_2(x) \geq \ldots \geq \hat{\lambda}_{r(\mathcal{P})}(x) \geq 0 \) and \( \hat{A}^\mathcal{P}(x) \) and \( \hat{\Gamma}^\mathcal{P}(x) \) are the \( \text{card} (\mathcal{P}) \times r(\mathcal{P}) \) matrices of left and right singular vectors, respectively. The diagonal elements of \( \hat{\Lambda}^\mathcal{P}(x) \) can be interpreted as the canonical correlations between \( x_\mathcal{P} \) and \( x_{\mathcal{P}^c} \), and the columns of \( (S_{\mathcal{P}, \mathcal{P}}(x))^{-\frac{1}{2}} \hat{A}^\mathcal{P}(x) \) and \( (S_{\mathcal{P}^c, \mathcal{P}^c}(x))^{-\frac{1}{2}} \hat{\Gamma}^\mathcal{P}(x) \) are the left and right canonical vectors.

Using the notation in (4), the LRT statistic in (3) can be rewritten as

\[
\frac{|S(x)|}{|S_{\mathcal{P}, \mathcal{P}}(x)||S_{\mathcal{P}^c, \mathcal{P}^c}(x)|} = \left| I_{r(\mathcal{P})} - \left( \hat{\Lambda}^\mathcal{P}(x) \right)^2 \right|,
\]

(5)

and the associated p-value can be rewritten as

\[
plrt(x; \mathcal{P}) = P_{H_0^\mathcal{P}} \left( \left| I_{r(\mathcal{P})} - \left( \hat{\Lambda}^\mathcal{P}(x) \right)^2 \right| \leq \left| I_{r(\mathcal{P})} - \left( \hat{\Lambda}^\mathcal{P}(x) \right)^2 \right| \right).
\]

(6)

Details are in Supplementary Material Section S1. Under \( H_0^\mathcal{P} \) and the assumption that \( n > p \), the test statistic follows a Wilks’ lambda distribution (see page 288 of Mardia et al. 1979). We will later use Proposition 1 to sample from the Wilks’ lambda distribution to evaluate \( plrt(x; \mathcal{P}) \) in (6).
Proposition 1. Under \( H_0^p \) and the assumption that \( n > p \), the following statements hold regarding the sample canonical correlations \( (\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_{r(P)}) \) defined in (4):

(i) The sample canonical correlations have joint density,

\[
 f \left( \hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_{r(P)} \right) = \pi^\frac{1}{2} \left( r(P) \right)^2 \frac{2^{r(P)} \Gamma(r(P)) \left( \frac{n}{2} \right)}{\Gamma(r(P)) \left( \frac{n-p+r(P)}{2} \right) \Gamma_p \left( \frac{1}{2} r \left( P \right) \right) \Gamma_p \left( \frac{1}{2} \left( p - r \left( P \right) \right) \right)} \times \\
 \prod_{i=1}^{r(P)} \left\{ \hat{\lambda}_i^{(p-2r(P))} \left( 1 - \hat{\lambda}_i^2 \right)^{\frac{1}{2} (n-p-2)} \right\} \prod_{i<j} \left( \hat{\lambda}_i^2 - \hat{\lambda}_j^2 \right),
\]

supported on \( \{ 1 \geq \hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_{r(P)} \geq 0 \} \).

(ii) The sample canonical correlations have the same distribution as

\[
 \sqrt{\frac{\Psi_1}{1+\Psi_1}} \sqrt{\frac{\Psi_2}{1+\Psi_2}} \ldots \sqrt{\frac{\Psi_{r(P)}}{1+\Psi_{r(P)}}}, \text{ where } (\Psi_1, \Psi_2, \ldots, \Psi_{r(P)}) \text{ are the eigenvalues of } W T^{-1},
\]

for \( W \sim \text{Wishart}_{r(P)}(I_{r(P)}, p - r(\mathcal{P})) \), \( T \sim \text{Wishart}_{r(P)}(I_{r(P)}, n - 1 - p + r(\mathcal{P})) \), and \( W \perp T \).

(iii) If \( r(\mathcal{P}) = 1 \), then the sample canonical correlation \( \hat{\lambda}_1 \) is the sample multiple correlation coefficient for the regression of the single variable onto the remaining \( p - 1 \) variables, and

\[
 \hat{\lambda}_1^2 \sim \text{Beta} \left( \frac{p-1}{2}, \frac{n-p}{2} \right).
\]

Here, (i) gives us the joint density of \( (\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_{r(P)}) \). In Section 4, we will use (ii) to sample from the joint distribution of \( (\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_{r(P)}) \), as the joint density in (i) is not amenable to sampling. We will make use of (iii) to bypass sampling in the special case of \( r(\mathcal{P}) = 1 \). We defer the proof of Proposition 1 to Supplementary Material Section S2. Crucially, under \( H_0^p \), the distribution of the sample canonical correlations depends only on \( n, p, \) and \( r(\mathcal{P}) \).

Next, we explore how the situation changes if we consider a data-dependent subset of variables, instead of a predetermined one as in the case of (2).

2.2 What changes when the groups of variables are chosen from the data?

We now suppose that \( \mathcal{P} \) is a function of the data, and specifically of \( S(x) \). That is, we define \( \mathcal{P} : S_{++}^p \rightarrow B_p \), where \( S_{++}^p \) denotes the set of all symmetric positive definite matrices, and \( B_p \) denotes the set of all possible partitions of \( \{1, 2, \ldots, p\} \). For \( \hat{\mathcal{P}} \in \mathcal{P}(S(x)) \), we wish to test

\[
 H_0^{\hat{\mathcal{P}}} : \Sigma_{\hat{\mathcal{P}}, \hat{\mathcal{P}}^c} = 0 \quad \text{versus} \quad H_1^{\hat{\mathcal{P}}} : \Sigma_{\hat{\mathcal{P}}, \hat{\mathcal{P}}^c} \neq 0.
\]
We assume without loss of generality that \( \hat{\mathcal{P}} = \{1, 2, \ldots, \text{card}(\hat{\mathcal{P}})\} \) and \( \hat{\mathcal{P}}^c = \{\text{card}(\hat{\mathcal{P}}) + 1, \ldots, p\} \). Since \( \hat{\mathcal{P}} \in \mathcal{P}(\mathbf{S}(\mathbf{x})) \), the null hypothesis in (9) is a function of the data. The classical test statistic in (3) does not follow a Wilks’ lambda distribution under the null hypothesis (9). Ignoring this fact can lead to a tremendous loss of power. To demonstrate this in an example, we simulate data with \( p = 6 \) and \( n = 9 \) for a completely dense \( \Sigma \), so that \( H_{0}^{\hat{\mathcal{P}}}: \Sigma_{\hat{\mathcal{P}}, \hat{\mathcal{P}}^c} \neq \mathbf{0} \) holds for all \( \hat{\mathcal{P}} \). Figures 1(a)-(b) display the heatmaps of the absolute values of the entries of the population correlation matrix and the sample correlation matrix. We tested (9) for \( \hat{\mathcal{P}} \) obtained via the thresholding procedure described in Algorithm 1; this yielded \( \hat{\mathcal{P}} = \{1, 2, 3\} \). Ignoring that \( \hat{\mathcal{P}} \) was selected based on the data and using Wilks’ lambda as the reference distribution for the likelihood ratio test statistic yielded a p-value \( (p_{LRT}) \) of 0.773, whereas our proposed selective inference approach produced a p-value of 0.015. Figure 1(c) shows the distribution of p-values over 10,000 data realizations using the two approaches. We see that our selective p-values tend to be small — as expected, since the null hypothesis does not hold — whereas the classical p-values are stochastically greater than a uniform.

At first glance, the fact that the classical test has low power in this scenario may seem counter-intuitive, since failure to account for selection typically leads to an inflated type I error rate, as opposed to low power, in related settings (Fithian et al., 2014). However, in this setting, the procedure \( \mathcal{P}(\cdot) \) that selects \( \hat{\mathcal{P}} \) (as described in Algorithm 1) does so because of its low correlation with \( \hat{\mathcal{P}}^c \). Thus, the likelihood ratio test statistic will be stochastically greater than the Wilks’ lambda distribution under \( H_{0}^{\hat{\mathcal{P}}}: \Sigma_{\hat{\mathcal{P}}, \hat{\mathcal{P}}^c} = \mathbf{0} \), yielding an overly conservative p-value.

**2.3 Selective inference for the covariance matrix**

Fithian et al. (2014) argues that when a null hypothesis is selected from the data, we should control the probability of a false rejection conditional on having selected this null hypothesis, i.e. the selective type I error rate. In the context of (9), the selective type I error rate is defined as follows:

**Definition 1.** (Selective type I error) A test of \( H_{0}^{R}: \Sigma_{R, R^c} = \mathbf{0} \) controls the selective
type I error if, for every subset $R$ of \{1, 2, \ldots, p\},

\[ \mathbb{P}_{H_0^R} \left( \text{reject } H_0^R \text{ at level } \alpha \big| R \in \mathcal{P}(S(X)) \right) \leq \alpha; \ \forall \ 0 \leq \alpha \leq 1. \]  

(10)

By the probability integral transform, it can be shown that a conditional version of (6),

\[ p_0 \left( x; \hat{P} \right) := \mathbb{P}_{H_0^P} \left( \left| I - \left( \hat{A}^P(X) \right)^2 \right| \leq \left| I - \left( \hat{A}^P(x) \right)^2 \right| \big| \hat{P} \in \mathcal{P}(S(X)) \right), \]

(11)

controls the selective type I error in (10). However, (11) is not computationally tractable, because the distribution of $\left| I - \left( \hat{A}^P(X) \right)^2 \right|$, conditional on $\hat{P} \in \mathcal{P}(S(X))$, depends on parameters that are unknown even under the null hypothesis. To bypass this problem, we employ a common approach in the selective inference literature (see e.g. Lee et al. 2016), in which we condition on some additional information. In particular, we propose a p-value for $H_0^P$ as

\[ p \left( x; \hat{P} \right) := \mathbb{P}_{H_0^P} \left( \left| I - \left( \hat{A}^P(X) \right)^2 \right| \leq \left| I - \left( \hat{A}^P(x) \right)^2 \right| \big| \hat{P} \in \mathcal{P}(S(X)) \right), \]

\[ S_{\hat{P},\hat{P}}(X) = S_{\hat{P},\hat{P}}(x), S_{\hat{P},\hat{P}}(X) = S_{\hat{P},\hat{P}}(x), \]

(12)

where we recall the SVD in (4). The intuition behind this conditioning set is as follows:

1. **Fixed within-group covariance:** Due to the nature of the null hypothesis in (9), we are primarily interested in inter-group correlation. Hence we keep the intra-group covariance fixed, i.e., we only consider $X$ for which $S_{\hat{P},\hat{P}}(X) = S_{\hat{P},\hat{P}}(x)$ and $S_{\hat{P},\hat{P}}(X) = S_{\hat{P},\hat{P}}(x)$.

2. **Fixed between-group distance direction:** Inspecting (4) and (5), we see that the only part of $S_{\hat{P},\hat{P}}^W$ that is involved in the test statistic is $\hat{A}^P$. The left and right singular vectors of $S_{\hat{P},\hat{P}}^W$ in (4), i.e. $\hat{A}^\ddagger$ and $\hat{\Gamma}^\ddagger$, are not involved in the test statistic. Hence, in (12), we only consider $X$ for which $\hat{A}^\ddagger(X) = \hat{A}^\ddagger(x)$ and $\hat{\Gamma}^\ddagger(X) = \hat{\Gamma}^\ddagger(x)$.

The next theorem establishes that the conditioning set in (12) can be written in terms of constraints on the sample canonical correlations, and that $p \left( x; \hat{P} \right)$ controls the selective type I error rate.

**Theorem 1.** For any realization $x$ of $X$ in (1) and for an arbitrary subset $R$ of \{1, 2, \ldots, p\}, we can express $p \left( x; \mathcal{R} \right)$ defined in (12) as

\[ p \left( x; \mathcal{R} \right) = \mathbb{P}_{H_0^R} \left( \prod_{i=1}^{r(\mathcal{R})} \left( 1 - \lambda_i^2 \right) \leq \prod_{i=1}^{r(\mathcal{R})} \left( 1 - \left( \hat{\lambda}_i^\ddagger(x) \right)^2 \right) \big| (\lambda_1, \lambda_2, \ldots, \lambda_{r(\mathcal{R})}) \in \mathcal{G}(x; \mathcal{R}) \right), \]

(13)
where \( \{\lambda_i\}_{i=1}^{r(\mathcal{R})} \) has joint density \( f(\lambda_1, \lambda_2, \ldots, \lambda_{r(\mathcal{R})}) \) defined in (7), \( \{\hat{\lambda}_i^\mathcal{R}(x)\}_{i=1}^{r(\mathcal{R})} \) are the diagonal entries of \( \hat{\mathbf{A}}^\mathcal{R}(x) \), and

\[
\mathcal{G}(x; \mathcal{R}) := \left\{ (\lambda_1, \lambda_2, \ldots, \lambda_{r(\mathcal{R})}) \in [0,1]^{r(\mathcal{R})} : \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{r(\mathcal{R})} ; \mathcal{R} \in \mathcal{P}(\mathbf{S}'(x, \mathcal{R}, \{\lambda_i\}_{i=1}^{r(\mathcal{R})})) \right\},
\]

where \( \mathbf{S}'(x, \mathcal{R}, \{\lambda_i\}_{i=1}^{r(\mathcal{R})}) \) is a modified version of the sample covariance matrix with a perturbed off-diagonal block,

\[
\mathbf{S}'_{\mathcal{R},\mathcal{R}^c}(x, \mathcal{R}, \{\lambda_i\}_{i=1}^{r(\mathcal{R})}) := (\mathbf{S}_{\mathcal{R},\mathcal{R}}(x))^\frac{1}{2} \hat{\mathbf{A}}^\mathcal{R}(x) \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{r(\mathcal{R})}) \left( \hat{\mathbf{G}}^\mathcal{R}(x) \right)^\top (\mathbf{S}_{\mathcal{R}^c,\mathcal{R}^c}(x))^\frac{1}{2}.
\]

Furthermore, rejecting \( H_0^\mathcal{R} \) if \( p(x; \mathcal{R}) \leq \alpha \) controls the selective type I error rate at level \( \alpha \), i.e.

\[
\mathbb{P}_{H_0^\mathcal{R}} \left( p(X; \mathcal{R}) \leq \alpha \bigg| \mathcal{R} \in \mathcal{P}(S(X)) \right) \leq \alpha; \ \forall \ 0 \leq \alpha \leq 1.
\]

Recall that \( \hat{\mathbf{A}}^\mathcal{R}(x) \), \( \hat{\mathbf{A}}_{\mathcal{R}^c}(x) \) and \( \hat{\mathbf{G}}^\mathcal{R}(x) \) were defined in (4). We defer the proof of Theorem 1 to Supplementary Material Section S3. Theorem 1 demonstrates that computation of the p-value in (12) boils down to characterizing \( \mathcal{G}(x; \hat{\mathbf{P}}) \) in (14).

Furthermore, \( \mathcal{G}(x; \hat{\mathbf{P}}) \) in (14) describes the points \( \{\lambda_i\}_{i=1}^{r(\hat{\mathbf{P}})} \) in the \( r(\hat{\mathbf{P}}) \)-dimensional unit hyper-cube for which the coordinates are in non-increasing order, and for which applying a variable grouping method \( \mathcal{P}(\cdot) \) to the perturbed covariance matrix \( \mathbf{S}'(x, \hat{\mathbf{P}}, \{\lambda_i\}_{i=1}^{r(\hat{\mathbf{P}})}) \) yields \( \hat{\mathbf{P}} \). Characterization of this set depends on the procedure used for obtaining the block diagonal structure of the correlation matrix, i.e. on \( \mathcal{P}(\cdot) \). In the next section, we focus on a specific data-adaptive procedure for \( \mathcal{P}(\cdot) \), and on the corresponding characterization of \( \mathcal{G}(x; \hat{\mathbf{P}}) \) in (14).

### 3 Characterization of the conditioning set

#### 3.1 Procedure for obtaining groups of uncorrelated variables

We present a simple procedure for identifying groups of uncorrelated variables, i.e., for discovering block diagonal structure of the covariance matrix. We assume that there are no ties between the off-diagonal entries of the sample correlation matrix, which holds with probability 1. We denote the correlation matrix corresponding to the sample covariance matrix \( \mathbf{S} = \mathbf{S}(x) \) as \( \mathbf{R}(\mathbf{S}) \), with \((i, j)^{th}\) element \( \mathbf{R}_{i,j}(\mathbf{S}) = \frac{S_{i,j}}{\sqrt{S_{i,i}S_{j,j}}} \), for \( i, j \in \{1, 2, \ldots, p\} \).

Let \( 1 \{ \mathcal{A} \} \) denote the indicator function of the event \( \mathcal{A} \). Algorithm 1 summarizes the procedure for obtaining groups of uncorrelated variables. Note that if the variables have an intrinsic ordering, then in line 5 of Algorithm 1 we discard groups of variables that violate that ordering.
Algorithm 1 Procedure for obtaining groups of uncorrelated variables

Input: Sample covariance matrix \( S(x) \); threshold \( c \).

1: procedure
2: Create the adjacency matrix \( D_{i,j} = 1 \{|R_{i,j}(S(x))| > c\} \);
3: Identify the connected components \( C_1, C_2, \ldots, C_K \) of the graph corresponding to the adjacency matrix;
4: if variables are ordered then
5: \( P(S(x)) = \{C_k : C_k \text{ consists exclusively of consecutive variables}\} \);
6: else
7: \( P(S(x)) = \{C_1, C_2, \ldots, C_K\} \).

3.2 Calculation of the conditioning set

In this subsection, we characterize the conditioning set \( G(x; PILE) \) in (14) for the function \( P(\cdot) \) defined in Algorithm 1.

Proposition 2. For \( P \in P(S(x)) \), we have that

\[
G(x; P) = \left\{ (\lambda_1, \lambda_2, \ldots, \lambda_r(P)) \in [0, 1]^r(P) : \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r(P), 
\right.
\]

\[
c \geq \max_{i' \in P, j' \in \overline{P}} \left| R_{i', j'} \left( S'(x, P, \{\lambda_i\}_{i=1}^r(P)) \right) \right| \}
\]

(17)

The proof is in Supplementary Material Section S4. The next proposition shows that the inter-group correlations appearing in the above expression are in fact linear combinations of \( \lambda_1, \lambda_2, \ldots, \lambda_r(P) \).

Proposition 3. The \( \text{card}(P) \times (p - \text{card}(P)) \) submatrix \( R_{\overline{P}, \overline{P}} \left( S'(x, P, \{\lambda_i\}_{i=1}^r(P)) \right) \)

is linear in \( (\lambda_1, \lambda_2, \ldots, \lambda_r(P)) \).

The proof of this proposition is provided in Supplementary Material Section S5. In the next theorem, we show that \( G(x; P) \) in (14) can be rewritten as a closed convex polytope in \( \mathbb{R}^r(P) \).

Theorem 2. For \( P \in P(S(x)) \) corresponding to the threshold \( c \) in Algorithm 1, there exist \( L = L(S(x), P) \in \mathbb{R}^{2\text{card}(P)(p - \text{card}(P)) + r(P) + 1} \times r(P) \) and \( g = g(P, c) \in \mathbb{R}^{2\text{card}(P)(p - \text{card}(P)) + r(P) + 1} \) such that

\[
G(x; P) = \left\{ \lambda = (\lambda_1, \lambda_2, \ldots, \lambda_r(P))^\top \in \mathbb{R}^r(P) : L\lambda \leq g \right\}.
\]

(18)

Thus, \( G(x; P) \) in (14) is a convex polytope. The proof of this theorem is provided in Supplementary Material Section S6.
4 Computation of $p(x; \hat{P})$ in (13)

We observe that $p(x; \hat{P})$ in (13) can be written as

$$p(x; \hat{P}) = \frac{\int_{\mathcal{L} \lambda \leq g} \left\{ \prod_{i=1}^{r} (1 - \lambda_i^2) \leq \prod_{i=1}^{r} (1 - \hat{\lambda}_i^2(x))^2 \right\} f \left( \lambda_1, \lambda_2, \ldots, \lambda_r(\hat{P}) \right) d\lambda_1 d\lambda_2 \ldots d\lambda_r(\hat{P})}{\int_{\mathcal{L} \lambda \leq g} \left\{ \prod_{i=1}^{r} (1 - \lambda_i^2) \leq \prod_{i=1}^{r} (1 - \hat{\lambda}_i^2(x))^2 \right\} f \left( \lambda_1, \lambda_2, \ldots, \lambda_r(\hat{P}) \right) d\lambda_1 d\lambda_2 \ldots d\lambda_r(\hat{P})},$$

where $f(\lambda_1, \lambda_2, \ldots, \lambda_r(\hat{P}))$, introduced in (7), is the joint density of the canonical correlations if $\hat{P}$ is prespecified, and the last equality follows from (18).

4.1 Simple computation when $r(\hat{P}) = 1$

In Supplementary Material Section S7.1, we show that if $r(\hat{P}) = 1$, then $p(x; \hat{P})$ in (19) can be written in terms of the cumulative distribution function of the univariate beta distribution:

**Proposition 4.** If $r(\hat{P}) = 1$, then there exists $g_u = g_u(S(x), \hat{P}, c) \in [0, 1]$ such that $p(x; \hat{P})$ in (19) can be written as function of truncated beta distributions, as follows:

$$p(x; \hat{P}) = \frac{F(g_u) - F \left( \min \left\{ g_u, \left( \hat{\lambda}_i^2(x) \right)^2 \right\} \right)}{F(g_u) - F(0)},$$

where $F$ is the cumulative distribution function of a Beta $\left( \frac{p-1}{2}, \frac{n-p}{2} \right)$ distribution.

In Proposition 4, $c$ is the threshold used in Algorithm 1.

4.2 Numerical integration when $r(\hat{P})$ is small

Evaluating (19) is challenging, because we do not have a closed form expression for the integrals involved. When $r(\hat{P})$ is a small number that exceeds one, we adopt methods of numerical integration on a convex polytope to evaluate (19). To evaluate the integral of a function over the convex polytope $\{ \lambda : L \lambda \leq g \}$, we first partition the polytope into simplices using Delaunay triangulation (Lee and Schachter, 1980), next integrate the corresponding function over those simplices, and finally sum those integrated values. Details of this computation to approximate (19) are provided in Supplementary Material Section S7.2.
4.3 Monte Carlo approximation when \( r(\hat{P}) \) is large

For moderately large values of \( r(\hat{P}) \), the approach in Section 4.2 is computationally taxing or infeasible. Thus, we resort to Monte Carlo approximation. From (19), we have that

\[
p(x; \hat{P}) = \frac{\mathbb{E}_{H_0^p} \left( \mathbb{I} \{ \prod_{i=1}^{r(\hat{P})} (1 - \lambda_i^2) \leq \prod_{i=1}^{r(\hat{P})} \left(1 - \left(\hat{\lambda}_i^p(x)\right)^2\right) , \ L\lambda \leq g \} \right)}{\mathbb{E}_{H_0^p} (\mathbb{I} \{ L\lambda \leq g \})}.
\]

(21)

For \( b = 1, 2, \ldots, B \), we simulate \( \lambda^{(b)} = (\lambda_1^{(b)}, \lambda_2^{(b)}, \ldots, \lambda_r^{(b)}) \) following (ii) in Proposition 1. Using (21), we then approximate \( p(x; \hat{P}) \) as

\[
\begin{align*}
p(x; \hat{P}) & \approx \frac{\sum_{b=1}^{B} \mathbb{I} \{ \prod_{i=1}^{r(\hat{P})} (1 - \lambda_i^{(b)})^2 \leq \prod_{i=1}^{r(\hat{P})} \left(1 - \left(\hat{\lambda}_i^p(x)\right)^2\right) , \ L\lambda^{(b)} \leq g \}}{\sum_{b=1}^{B} \mathbb{I} \{ L\lambda^{(b)} \leq g \}}.
\end{align*}
\]

(22)

The overall procedure for computing \( p(x; \hat{P}) \) is summarized in Algorithm S1 of Supplementary Material Section S7.3.

5 Simulation results

5.1 Type I error under global null

We simulate data with unordered variables from (1) with \( \Sigma = I \), so that \( H_0^p \) holds for all partitions of the variables. We fix \( p = 100 \) and vary \( n \in \{1.1p, 1.5p, 2p\} \). For each simulated data set, we compute both the classical p-value \( p_{LRT} \) in (6) and the selective p-value in (12) for the hypothesis \( H_0^p : \Sigma_{\hat{P}, \hat{P}c} = 0 \) for a randomly chosen \( \hat{P} \in \mathcal{P}(\mathcal{S}(x)) \) with the procedure \( \mathcal{P}(\cdot) \) defined in Section 3.1 with the threshold \( c = 0.2 \). We approximate the selective p-value in (12) as discussed in Algorithm S1 of Supplementary Material Section S7.3.

Figure 2 displays the QQ plots of the empirical distribution of the p-values against the Uniform(0, 1) distribution, over 100,000 simulated data set. The classical p-value is overly conservative, whereas the proposed selective p-value is well-calibrated.
Figure 2: For 100,000 draws from (1), with $\Sigma = I_{100}$, $n \in \{1.1p, 1.5p, 2p\}$, QQ-plots are displayed for: (a) the “classical” p-value in (6); (b) the “selective” p-value in (12), approximated as described in Section 4.

5.2 Power at $\alpha = 0.05$

We simulate data with unordered variables from (1), with $p = 100$, $n \in \{1.1p, 1.5p, 2p\}$. For each simulated data set, we generate a random $p \times p$ positive definite matrix $\Sigma$ and a corresponding threshold $c = c(\Sigma)$ as described in Supplementary Material Section S8, and we randomly select $\hat{P} \in \mathcal{P}(S(x))$ with $\mathcal{P}(\cdot)$ defined in Section 3.1. We test $H_0^\hat{P}: \Sigma_{\hat{P}, \hat{P}c} = 0$ at significance level $\alpha = 0.05$. Motivated by the test statistic in (3), we define the effect size based on its population counterpart as $\Delta := 1 - |\Sigma| / |\Sigma_{\hat{P}, \hat{P}c}|$, and consider the probability of rejecting $H_0^\hat{P}: \Sigma_{\hat{P}, \hat{P}c} = 0$ as a function of $\Delta$ over 1,000,000 simulated data sets. Figure 3 shows that the proposed selective inference approach has higher power than classical inference for all values of $\Delta$ and $n$.

6 Application to gene co-expression networks

A gene co-expression network is an undirected graph in which the nodes represent genes and the edges represent pairs of genes that are co-expressed, in the sense of e.g. Pearson correlation (Guttman et al., 2011). Although gene co-expression networks do not confer information about causality, they are frequently used as a starting point for investigating gene regulation, functional enrichment, and hub genes (Van Dam et al., 2018). Addition-
Figure 3: Power estimates as a function of the effect-size $\Delta$ for 1,000,000 draws from the simulation setup described in Section 5.2, when (a) $r(\hat{P}) \leq 5$, (b) $r(\hat{P}) > 5$, where $r(\hat{P}) = \min\{\text{card}(\hat{P}), \text{card}(\hat{P}^c)\}$. The x-axis displays bins of $\Delta$ of size 0.1; only bins with more than 100 simulated datasets are displayed.

ally, they have proven useful in identifying disease genes through the guilt-by-association principle, enabling a better understanding of disease origin (Alsina et al., 2014) and progression (Chaussabel et al., 2008). For a detailed review of gene co-expression networks, we refer readers to Saelens et al. (2018) and references therein.

In this section, we focus on the gene expression data corresponding to *E. coli* and *S. cerevisiae* from the DREAM5 network inference challenge (Marbach et al., 2012). We will test for correlation between sets of genes that are known to be correlated based on outside biological knowledge. Our results indicate that our proposed selective inference approach has higher power than the classical inference approach.

### 6.1 *Escherichia coli* and *Saccharomyces cerevisiae* data sets

The DREAM5 *E. coli* data set consists of pre-processed gene expression measurements for $p = 4,297$ genes on $n = 805$ microarray chips (Marbach et al., 2012). Furthermore, the *E. coli* gene co-expression network is very well understood, and known edges can be found in publicly available databases, e.g. RegulonDB (Gama-Castro et al., 2010) and EcoCyc (Keseler et al., 2010). As part of the DREAM5 challenge, a highly curated co-expression network, based on this outside biological information, was made available. We treat this
as a “ground truth” co-expression network in what follows.

The DREAM5 *S. cerevisiae* data set consists of \( p = 5,667 \) pre-processed gene expression measurements on \( n = 536 \) microarrays. While the *S. cerevisiae* network is not as well understood as the *E. coli* network, the DREAM5 organizers combined data from 16 sources to arrive at a single co-expression network (MacIsaac et al., 2006). We treat this network as the “ground truth” in what follows.

### 6.2 Power to detect correlation between sets of genes

For each data set, we identify the gene with the most edges in the ground truth co-expression network, and restrict our analysis to that hub gene and 99 genes randomly sampled from the genes to which it is connected. This yields a set of \( p = 100 \) genes that form a single connected component in the ground truth co-expression network. By construction, for any \( \mathcal{P} \in \{1, 2, \ldots, p\} \), the null hypothesis \( H_0^P : \Sigma_{\hat{\mathcal{P}}, \hat{\mathcal{P}}} = 0 \) will not hold.

Next, we sample \( n \in \{1.2p, 1.4p, 1.5p, 2p\} \) microarrays, and let \( \mathbf{x} \) denote the resulting \( n \times p \) data set. We then compute \( \mathcal{P}(\mathbf{S}(\mathbf{x})) \) using the procedure in Section 3.1. To choose the threshold \( c \), we apply the procedure described in Supplementary Material Section S8 to the sample covariance matrix computed on the held-out microarrays (which we write as \( \hat{\mathbf{S}}_{\text{holdout}} \)), i.e. those not included in \( \mathbf{x} \). Since the microarrays are independent, this threshold \( c \) is independent of \( \mathbf{x} \). For a randomly selected \( \hat{\mathcal{P}} \in \mathcal{P}(\mathbf{S}(\mathbf{x})) \), we test \( H_0^\hat{\mathcal{P}} : \Sigma_{\hat{\mathcal{P}}, \hat{\mathcal{P}}} = 0 \) at significance level \( \alpha = 0.05 \), where \( \Sigma \) denotes the population covariance matrix of the \( p \) genes. We repeat this entire procedure 10,000 times for different random sets of 99 neighbors of the hub gene.

Figures 4(a) and 5(a) display the probability of rejecting \( H_0^\hat{\mathcal{P}} : \Sigma_{\hat{\mathcal{P}}, \hat{\mathcal{P}}} = 0 \) as a function of \( \hat{\Delta} := 1 - \frac{|\hat{\mathbf{S}}_{\text{holdout}}|}{|\hat{\mathbf{S}}_{\text{holdout}, \mathcal{P}}| |\hat{\mathbf{S}}_{\text{holdout}, \mathcal{P}, \mathcal{P}}|} \). In this setting, by construction, the \( p = 100 \) genes have extremely high correlation. For any \( \hat{\mathcal{P}} \in \mathcal{P}(\mathbf{S}(\mathbf{x})) \), the conditioning set in (13) deals with a modified version \( \mathbf{S}'(\mathbf{x}, \hat{\mathcal{P}}, \cdot) \) of the sample covariance matrix \( \mathbf{S}(\mathbf{x}) \) with a perturbed off-diagonal block (recall Theorem 1). Because the genes are highly correlated, for perturbations under \( H_0^\hat{\mathcal{P}} \), with very high probability, we will have \( \mathcal{P}\left(\mathbf{S}'(\mathbf{x}, \hat{\mathcal{P}}, \cdot)\right) = \mathcal{P}(\mathbf{S}(\mathbf{x})) \). Hence the classical and selective approaches yield almost identical results.

To investigate the performance of the classical and selective approaches when the correlation between genes is lower, we add Gaussian white noise to each gene’s expression, as follows:

\[
\tilde{\mathbf{x}}_j = \mathbf{x}_j + \eta \varepsilon_j, \quad \varepsilon_j \sim N(0, \hat{\var} \mathbf{x}_j) \mathbf{I}, \quad j = 1, 2, \ldots, p.
\]

Here \( \mathbf{x}_j \in \mathbb{R}^n \) is the expression vector for the \( j \)th gene, i.e. the \( j \)th column of \( \mathbf{x} \). We vary the noise level \( \eta \in \{0.5, 3\} \). In Figures 4 and 5, we see that the power of both the classical and selective approaches decreases as \( \eta \) increases. Moreover, as \( \eta \) increases, the correlation between the genes decreases. Hence for perturbations under \( H_0^\hat{\mathcal{P}} \), the probability of the event \( \mathcal{P}\left(\mathbf{S}'(\mathbf{x}, \hat{\mathcal{P}}, \cdot)\right) = \mathcal{P}(\mathbf{S}(\mathbf{x})) \) decreases. As a result, the difference between the classical and selective approaches increases, and the selective approach becomes more powerful than the classical approach.
Figure 4: Power estimates as a function of $\hat{\Delta}$ for *E. coli* for three values of $\eta$ in (23). (a) $\eta = 0$, (b) $\eta = 0.5$, (c) $\eta = 3$ in (23). We only display bins of $\hat{\Delta}$ with more than 100 observations.

Figure 5: Power estimates as a function of $\hat{\Delta}$ for *S. cerevisiae* for three values of $\eta$ in (23). (a) $\eta = 0$, (b) $\eta = 0.5$, (c) $\eta = 3$ in (23). We only display bins of $\hat{\Delta}$ with more than 100 observations.

7 Discussion

Our proposed approach for testing dependence between data-driven groups of Gaussian variables requires the number of features to be less than the number of observations, i.e. $p < n$, in order for the test statistic to be defined. Future work could combine tools from high-dimensional inference with selective inference to account for selection in covariance matrices with $p \gg n$.

The normality assumption on the variables is needed to obtain Proposition 1, but is restrictive in practice. For example, RNA sequencing data involve nonnegative counts, which are better modeled with a Poisson (Witten, 2011) or negative binomial (Risso et al., 2018) distribution. Extending the proposed selective inference approach to accommodate a more general class of distributions requires further investigation.

An R package to implement the proposed approach called `independencepvalue` is available at https://arkajyotisaha.github.io/independencepvalue-project.
References

Ali, H. and Nagar, D. K. (2002). Null distribution of multiple correlation coefficient under mixture normal model. International Journal of Mathematics and Mathematical Sciences, 30(4):249–255.

Alsina, L., Israelsson, E., Altman, M. C., Dang, K. K., Ghandil, P., Israel, L., Von Bernuth, H., Baldwin, N., Qin, H., Jin, Z., et al. (2014). A narrow repertoire of transcriptional modules responsive to pyogenic bacteria is impaired in patients carrying loss-of-function mutations in MYD88 or IRAK4. Nature Immunology, 15(12):1134–1142.

Anderson, T. W. (2003). An introduction to multivariate statistical analysis. Wiley, 3rd edition.

Ben-Israel, A. (1999). The change-of-variables formula using matrix volume. SIAM Journal on Matrix Analysis and Applications, 21(1):300–312.

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. Annals of Statistics, 36(6):2577–2604.

Bien, J. and Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. Biometrika, 98(4):807–820.

Bordier, C., Nicolini, C., and Bifone, A. (2017). Graph analysis and modularity of brain functional connectivity networks: searching for the optimal threshold. Frontiers in Neuroscience, 11:441.

Cai, T. T. and Yuan, M. (2012). Adaptive covariance matrix estimation through block thresholding. Annals of Statistics, 40(4):2014–2042.

Chaussabel, D., Quinn, C., Shen, J., Patel, P., Glaser, C., Baldwin, N., Stichweh, D., Blankenship, D., Li, L., Munagala, I., et al. (2008). A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. Immunity, 29(1):150–164.

Chen, S. and Bien, J. (2020). Valid inference corrected for outlier removal. Journal of Computational and Graphical Statistics, 29(2):323–334.

Chen, Y. T. and Witten, D. M. (2022). Selective inference for k-means clustering. arXiv preprint arXiv:2203.15267.

Drton, M. and Perlman, M. D. (2008). A SINful approach to gaussian graphical model selection. Journal of Statistical Planning and Inference, 138(4):1179–1200.

Fithian, W., Sun, D., and Taylor, J. (2014). Optimal inference after model selection. arXiv preprint arXiv:1410.2597.

Freeman, T. C., Goldovsky, L., Brosch, M., Van Dongen, S., Mazière, P., Grocock, R. J., Freilich, S., Thornton, J., and Enright, A. J. (2007). Construction, visualisation, and clustering of transcription networks from microarray expression data. PLoS Computational Biology, 3(10):e206.
Gama-Castro, S., Salgado, H., Peralta-Gil, M., Santos-Zavaleta, A., Muniz-Rascado, L., Solano-Lira, H., Jimenez-Jacinto, V., Weiss, V., Garcia-Sotelo, J. S., Lopez-Fuentes, A., et al. (2010). RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). Nucleic Acids Research, 39(suppl_1):D98–D105.

Gao, L. L., Bien, J., and Witten, D. (2022). Selective inference for hierarchical clustering. Journal of the American Statistical Association (To appear in).

Geyer, C. J., Meeden, G. D., and incorporates code from cddlib written by Komei Fukuda (2021). rcdd: Computational Geometry. R package version 1.5.

Guttman, M., Donaghey, J., Carey, B. W., Garber, M., Grenier, J. K., Munson, G., Young, G., Lucas, A. B., Ach, R., Bruhn, L., et al. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature, 477(7364):295–300.

Habel, K., Grasman, R., Gramacy, R. B., Mozharovskyi, P., and Sterratt, D. C. (2019). geometry: Mesh Generation and Surface Tessellation. R package version 0.4.5.

Hsu, P. (1939). On the distribution of roots of certain determinantal equations. Annals of Eugenics, 9(3):250–258.

Hyun, S., G’sell, M., and Tibshirani, R. J. (2018). Exact post-selection inference for the generalized lasso path. Electronic Journal of Statistics, 12(1):1053–1097.

Jendoubi, T. and Strimmer, K. (2019). A whitening approach to probabilistic canonical correlation analysis for omics data integration. BMC Bioinformatics, 20(1):1–13.

Jewell, S., Fearnhead, P., and Witten, D. (2022). Testing for a change in mean after changepoint detection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) (To appear in).

Keseler, I. M., Collado-Vides, J., Santos-Zavaleta, A., Peralta-Gil, M., Gama-Castro, S., Muniz-Rascado, L., Bonavides-Martinez, C., Paley, S., Krummenacker, M., Altman, T., et al. (2010). EcoCyc: a comprehensive database of Escherichia coli biology. Nucleic Acids Research, 39(suppl_1):D583–D590.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. Annals of Statistics, 37(6B):4254.

Lee, D.-T. and Schachter, B. J. (1980). Two algorithms for constructing a Delaunay triangulation. International Journal of Computer & Information Sciences, 9(3):219–242.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. Annals of Statistics, 44(3):907–927.

Lee, J. D. and Taylor, J. E. (2014). Exact post model selection inference for marginal screening. arXiv preprint arXiv:1402.5596.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. Annals of Statistics, 40(4):2293–2326.
Liu, H., Wang, L., and Zhao, T. (2014). Sparse covariance matrix estimation with eigenvalue constraints. Journal of Computational and Graphical Statistics, 23(2):439–459.

Loftus, J. R. and Taylor, J. E. (2015). Selective inference in regression models with groups of variables. arXiv preprint arXiv:1511.01478.

MacIsaac, K. D., Wang, T., Gordon, D. B., Gifford, D. K., Stormo, G. D., and Fraenkel, E. (2006). An improved map of conserved regulatory sites for Saccharomyces cerevisiae. BMC Bioinformatics, 7(1):1–14.

Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., Kellis, M., Collins, J. J., and Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. Nature Methods, 9(8):796–804.

Mardia, K., Kent, J., and Bibby, J. (1979). Multivariate Analysis. Probability and Mathematical Statistics. Academic Press, 1st edition.

Neufeld, A. C., Gao, L. L., and Witten, D. M. (2021). Tree-values: selective inference for regression trees. arXiv preprint arXiv:2106.07816.

Nolan, J. P., with parts adapted from Fortran, and code by Alan Genz, M. (2021). Simplicial Cubature: Integration of Functions Over Simplices. R package version 1.3.7.

Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C. R., Lee, D. H., Marjoribanks, C., McDonough, D. P., et al. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science, 294(5547):1719–1723.

Qiu, W. and Joe, H. (2020). clusterGeneration: Random Cluster Generation with Specified Degree of Separation. R package version 1.3.7.

Rennie, J. (2006). Jacobian of the singular value decomposition with application to the trace norm distribution. http://qwone.com/~jason/writing/svdJacobian.pdf, 60.

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.-P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. Nature Communications, 9(1):1–17.

Rothman, A. J. (2012). Positive definite estimators of large covariance matrices. Biometrika, 99(3):733–740.

Rothman, A. J., Levina, E., and Zhu, J. (2009). Generalized thresholding of large covariance matrices. Journal of the American Statistical Association, 104(485):177–186.

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

Serra, A., Coretto, P., Fratello, M., and Tagliaferri, R. (2018). Robust and sparse correlation matrix estimation for the analysis of high-dimensional genomics data. Bioinformatics, 34(4):625–634.
Suzumura, S., Nakagawa, K., Umezu, Y., Tsuda, K., and Takeuchi, I. (2017). Selective inference for sparse high-order interaction models. In International Conference on Machine Learning, pages 3338–3347. PMLR.

Tian, X., Loftus, J. R., and Taylor, J. E. (2018). Selective inference with unknown variance via the square-root lasso. Biometrika, 105(4):755–768.

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. Journal of the American Statistical Association, 111(514):600–620.

Van Dam, S., Vosa, U., van der Graaf, A., Franke, L., and de Magalhaes, J. P. (2018). Gene co-expression analysis for functional classification and gene–disease predictions. Briefings in Bioinformatics, 19(4):575–592.

Witten, D. M. (2011). Classification and clustering of sequencing data using a Poisson model. The Annals of Applied Statistics, 5(4):2493–2518.

Xue, L., Ma, S., and Zou, H. (2012). Positive-definite \( l_1 \)-penalized estimation of large covariance matrices. Journal of the American Statistical Association, 107(500):1480–1491.

Yang, F., Foygel Barber, R., Jain, P., and Lafferty, J. (2016). Selective inference for group-sparse linear models. Advances in Neural Information Processing Systems, 29:2469–2477.

Zhao, T., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in R. The Journal of Machine Learning Research, 13(1):1059–1062.
Supplementary Materials for “Inferring independent sets of Gaussian variables after thresholding correlations"

S1  Brief review of canonical correlation analysis

Here we briefly review canonical correlation analysis, as it paves the way for the selective inference method proposed in this paper. We largely follow the notation used in Anderson (2003).

We partition the random vector \( X \sim N(0, \Sigma) \) of length \( p \) into two vectors \( X_1 \) and \( X_2 \) of lengths \( p_1 \) and \( p_2 = p - p_1 \) respectively, so that \( X^\top = (X_1^\top \; X_2^\top) \). The corresponding covariance matrix can be partitioned as

\[
\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},
\]

where \( \Sigma \) is of dimension \( p \times p \), \( \Sigma_{12} \) is of dimension \( p_1 \times p_2 \), etc. Without loss of generality, we assume \( p_1 \leq p_2 \). We define the population canonical correlations \( \lambda_1, \lambda_2, \ldots, \lambda_{p_1} \) between \( X_1 \) and \( X_2 \), and their corresponding population canonical vectors \( A^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_{p_1}^*) \in \mathbb{R}^{p_1 \times p_1} \) and \( \Gamma^* = (\gamma_1^*, \gamma_2^*, \ldots, \gamma_{p_1}^*) \in \mathbb{R}^{p_2 \times p_1} \), sequentially for \( r = 1, 2, \ldots, p_1 \):

\[
(\alpha_r^*, \gamma_r^*) := \arg \max_{\alpha_r^\top \Sigma_{11} \alpha_r = 1, \gamma_r^\top \Sigma_{22} \gamma_r = 1} \alpha_r^\top \Sigma_{12} \gamma_r; \quad \lambda_r := (\alpha_r^*)^\top \Sigma_{12} \gamma_r^*.
\]

The above can be solved through an SVD on \( \Sigma_{11}^{-\frac{1}{2}} \Sigma_{12} \Sigma_{22}^{-\frac{1}{2}} = A \Lambda \Gamma^\top \), by taking \( A^* = \Sigma_{11}^{-\frac{1}{2}} A, \Gamma^* = \Sigma_{22}^{-\frac{1}{2}} \Gamma \), and \( \Lambda_r = \text{diag} (\Lambda) \) (Jendoubi and Strimmer, 2019). Next, we define the sample canonical correlations and sample canonical vectors for a realization \( x \in \mathbb{R}^{n \times p} \) of \( X \) in (1). We assume \( \text{rank} (x) = p \), which holds with probability 1, since we assume \( \Sigma \) is positive definite. We let \( S = \begin{pmatrix} S_{11} & S_{12} \\ S_{12} & S_{22} \end{pmatrix} \) denote the sample covariance of \( x = (x_1^\top \; x_2^\top)^\top \), suitably partitioned. We now define the “whitened” versions of \( x_1 \) and \( x_2 \), i.e. \( x_1^W = S_{11}^{-\frac{1}{2}} x_1 \) and \( x_2^W = S_{22}^{-\frac{1}{2}} x_2 \). Let \( S^W = \begin{pmatrix} S_{11}^W & S_{12}^W \\ S_{12}^W & S_{22}^W \end{pmatrix} \) denote the sample covariance matrix of \( x^W = \begin{pmatrix} (x_1^W)^\top \\ (x_2^W)^\top \end{pmatrix}^\top \), suitably partitioned. It follows that \( S_{11}^W = I_{p_1}; S_{22}^W = I_{p_2}, \) and \( S_{12}^W = S_{11}^{-\frac{1}{2}} S_{12} S_{22}^{-\frac{1}{2}} \). The compact SVD of \( S_{12}^W \) is

\[
S_{12}^W \text{ SVD } \tilde{A} A \hat{\Gamma}^\top,
\]

where \( \tilde{A} \) is a square diagonal matrix of \( p_1 \) positive singular values, arranged in non-increasing order, and \( A \) and \( \hat{\Gamma} \) are the \( p_1 \times p_1 \) and \( p_2 \times p_1 \) matrices of left and right singular vectors, respectively. We now derive (5). We can rewrite the classical LRT statistic in (3)
by noting that

$$|S| = |S_{22}| |S_{11} - S_{12}S_{22}^{-1}S_{21}|$$

$$= |S_{22}| |S_{11}| \left| I - S_{11}^{-\frac{1}{2}}S_{12}S_{22}^{-\frac{1}{2}}S_{21}S_{11}^{-\frac{1}{2}} \right|$$

$$= |S_{11}| |S_{22}| \left| I - S_{12}^W (S_{12}^W)^\top \right|$$

$$= |S_{11}| |S_{22}| \left| I - \hat{A}^2 \right|,$$

where in the last equality, we have used that $\hat{A}$ is a square orthogonal matrix.

**S2  Proof of Proposition 1**

For any subset $\mathcal{P}$ of $\{1, 2, \ldots, p\}$, the sample canonical correlations are the diagonal elements of $\hat{A}^{\mathcal{P}}$, denoted by $\{\hat{\lambda}_i^{\mathcal{P}}\}_{i=1}^{r(\mathcal{P})}$. In this section, we will omit the $\mathcal{P}$ superscript and write $\{\hat{\lambda}_i\}_{i=1}^{r(\mathcal{P})}$.

**S2.1 Proof of Proposition 1(i)**

*Proof.* The joint density of $\left(\hat{\lambda}_1^2, \hat{\lambda}_2^2, \ldots, \hat{\lambda}_{r(\mathcal{P})}^2\right)$ under $H_0^\mathcal{P}$ takes the form

$$f\left(\hat{\lambda}_1^2 = \gamma_1, \hat{\lambda}_2^2 = \gamma_2, \ldots, \hat{\lambda}_{r(\mathcal{P})}^2 = \gamma_{r(\mathcal{P})}\right) = \pi^{\frac{1}{2} r(\mathcal{P})} \frac{\Gamma_r(\frac{r}{2})}{\Gamma_r(\frac{n-p+r(\mathcal{P})}{2}) \Gamma_p\left(\frac{1}{2}r(\mathcal{P})\right) \Gamma_p\left(\frac{1}{2}(p-r(\mathcal{P}))\right)} \prod_{i=1}^{r(\mathcal{P})} \left\{ \gamma_i \left(1 - \gamma_i\right)^{\frac{1}{2}(n-p-2)} \right\} \prod_{i<j} \left(\hat{\lambda}_i - \hat{\lambda}_j\right)$$

(S2)

by (13) in Section 13.4 in Anderson (2003). To obtain $f\left(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_{r(\mathcal{P})}\right)$ in (7), we perform a change of variables. We have

$$f\left(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_{r(\mathcal{P})}\right) = f\left(\hat{\lambda}_1^2, \hat{\lambda}_2^2, \ldots, \hat{\lambda}_{r(\mathcal{P})}^2\right) |J|,$$

(S3)

where $J$ is the $r(\mathcal{P}) \times r(\mathcal{P})$ Jacobian matrix of the transformation, with $J_{i,j} = \frac{\partial \hat{\lambda}_j^2}{\partial \hat{\lambda}_i}$ for $i,j \in \{1, 2, \ldots, r(\mathcal{P})\}$. Simplifying, we have $|J| = 2^{r(\mathcal{P})} \prod_{i=1}^{r(\mathcal{P})} \hat{\lambda}_i$. Substituting this into (S3) yields

$$f\left(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_{r(\mathcal{P})}\right) = \pi^{\frac{1}{2} r(\mathcal{P})} \frac{\Gamma_r(\frac{r}{2})}{\Gamma_r(\frac{n-p+r(\mathcal{P})}{2}) \Gamma_p\left(\frac{1}{2}r(\mathcal{P})\right) \Gamma_p\left(\frac{1}{2}(p-r(\mathcal{P}))\right)} \prod_{i=1}^{r(\mathcal{P})} \left\{ \hat{\lambda}_i \left(1 - \hat{\lambda}_i^2\right)^{\frac{1}{2}(n-p-2)} \right\} \prod_{i<j} \left(\hat{\lambda}_i^2 - \hat{\lambda}_j^2\right) \times 2^{r(\mathcal{P})} \prod_{i=1}^{r(\mathcal{P})} \hat{\lambda}_i,$$

(S4)

which completes the proof of (i) of Proposition 1. □
S2.2 Proof of Proposition 1(ii)

Proof. The result follows by applying various arguments from Anderson (2003). For clarity, we extract the relevant parts here. For notational ease, we will write $S(x)$ as $S$.

1) Applying the argument in (48)-(52) of Section 12.2 of Anderson (2003) to the sample covariance matrix, we have that the squared canonical correlations $\left\{ \hat{\lambda}_i^2 \right\}_{i=1}^{r(p)}$ are the roots, with respect to $\gamma$, of

$$f(\gamma) = \left| S_{p,p} (S_{p^e,p^e})^{-1} S_{p^e,p} - \gamma S_{p,p} \right|. \quad (S5)$$

2) By Corollary 4.3.2 of Anderson (2003), if $\Sigma_{p,p^e} = 0$, then

$$W := n (\Sigma_{p,p})^{-\frac{1}{2}} S_{p,p^e} (S_{p^e,p^e})^{-1} S_{p^e,p} (\Sigma_{p,p})^{-\frac{1}{2}} \sim \text{Wishart}_{r(p)} (I, p - r(P)), \quad T := n (\Sigma_{p,p})^{-\frac{1}{2}} (S_{p^e,p} - S_{p,p^e} (S_{p^e,p^e})^{-1} S_{p^e,p}) (\Sigma_{p,p})^{-\frac{1}{2}} \sim \text{Wishart}_{r(p)} (I, n - 1 + p - r(P)), \quad \text{and } W \text{ and } T \text{ are independent.}$$

3) Equations 7-10 of Section 13.2 of Anderson (2003) establish that the values of $\gamma$ that satisfy $|W - \gamma(T + W)| = 0$ are given by $\frac{\Psi_i(WT^{-1})}{1 + \Psi_i(WT^{-1})}$, where $\Psi_i(WT^{-1})$ are the eigenvalues of $WT^{-1}$.

By 1), $\hat{\lambda}_1^2, \hat{\lambda}_2^2, \ldots, \hat{\lambda}_{r(p)}^2$ are given by roots of $(S5)$. Pre- and post-multiplying $S_{p,p^e} (S_{p^e,p^e})^{-1} S_{p^e,p} - \gamma S_{p,p}$ in $(S5)$ by $\sqrt{n}\Sigma_{p,p}^{-\frac{1}{2}}$ gives that $\hat{\lambda}_1^2, \hat{\lambda}_2^2, \ldots, \hat{\lambda}_{r(p)}^2$ are the values of $\gamma$ that solve

$$|W - \gamma(T + W)| = 0,$$

where $W$ and $T$ are defined in 2). Applying 3) completes the proof of (ii) of Proposition 1. \qed

S2.3 Proof of Proposition 1(iii)

Proof. To prove Proposition 1(iii), we start with a result whose proof follows from (5) in Section 4.4.1 of Anderson (2003).

Proposition 5. The squared sample multiple correlation coefficient for regression for a response $h$ on a set of variables $Q$ is given by $s_{Q,h}(\Sigma_{Q,Q})^{-\frac{1}{2}}s_{Q,h}$, where $s_{Q,h}$ denotes the vector of covariances between $h$ and the variables in $Q$, and $\Sigma_{Q,Q}$ denotes the covariance matrix of the variables in $Q$.

Without loss of generality, we assume $\text{card}(P) = p - 1$. First we establish that the canonical correlation is the multiple correlation coefficient between the variable in $P^e$ and the remaining $p - 1$ variables in $P$. We write $S(x)$ as $S$ and denote the vectors $S_{p,p^e}^W$ and $S_{p,p^e}$ as $s_{p,p^e}^W$ and $s_{p,p^e}$, respectively, and the scalar $S_{p^e,p^e}$ as $s_{p^e,p^e}$. Substituting these in (4) we have

$$s_{p^e,p^e}^W = \frac{(S_{p,p})^{-\frac{1}{2}} s_{p,p^e}}{\sqrt{s_{p^e,p^e}}}.$$
Using Proposition 5, squared sample multiple correlation coefficient follows a beta distribution with parameters.

Recalling the definition in (15), we see that

\[ (\hat{\lambda}_1)^2 = \|s_{W, p, c}\|^2 = (s_{W, p, c}^T s_{W, p, c}) = \frac{s_{p, p, c}^T (s_{p, p})^{-1} s_{p, p, c}}{s_{p, p, c}}. \]

Using Proposition 5, \( s_{p, p, c}^T (s_{p, p})^{-1} s_{p, p, c} \) is the squared multiple correlation coefficient for the regression of the single variable in \( p^c \) onto the remaining \( p - 1 \) variables. Under the null, the squared sample multiple correlation coefficient follows a beta distribution with parameters \( \frac{p - 1}{2} \) and \( \frac{n - p}{2} \) (Ali and Nagar, 2002). This completes the proof of (iii) of Proposition 1. 

**S3 Proof of Theorem 1**

In this section, we abbreviate \( H_0^R, \hat{\Lambda}^R (\cdot), \hat{\Lambda}^R (\cdot), \) and \( \hat{\Gamma}^R (\cdot) \) (defined in Section 2) by omitting the superscript. Without loss of generality, we assume that \( R = \{1, 2, \ldots, p_1\}, \)

\( \mathcal{R}^c = \{p_1 + 1, \ldots, p\}, \) and \( p_1 \leq p/2. \) Define \( p_2 = p - p_1. \) Let \( \hat{\lambda}_1 (X), \hat{\lambda}_2 (X), \ldots, \hat{\lambda}_{p_1} (X) \)

be the diagonal elements of \( \hat{\Lambda} (X). \)

We let \( S (X) = \begin{pmatrix} S_{11} (X) & S_{12} (X) \\ S_{12}^T (X) & S_{22} (X) \end{pmatrix} \)

denote the covariance matrix of \( X = (X_1 \ X_2), \)

suitably partitioned, where \( X_1 \) and \( X_2 \) are defined as the submatrices of \( X \) with columns in \( R \) and \( R^c, \) respectively. Equipped with this notation, we first state a technical lemma whose proof we defer to Section S3.1.

**Lemma 1.** Under \( H_0 : \Lambda = 0, \hat{\Lambda} (X) \) is independent of \((S_{11} (X), S_{22} (X), \hat{\Lambda} (X), \hat{\Gamma} (X)). \)

By (4), \( \hat{\Lambda} (X) \hat{\Lambda} (X) \left( \hat{\Gamma} (X) \right)^T \)

is the SVD of \((S_{11} (X))^{-\frac{1}{2}} S_{12} (X) (S_{22} (X))^{-\frac{1}{2}} \) and so

\[ S_{12} (X) = (S_{11} (X))^{\frac{1}{2}} \hat{\Lambda} (X) \hat{\Lambda} (X) \left( \hat{\Gamma} (X) \right)^T (S_{22} (X))^{\frac{1}{2}}. \]

In (12), we condition on the additional information

\[ \left\{ S_{11} (X) = S_{11} (x), S_{22} (X) = S_{22} (x), \hat{\Lambda} (X) = \hat{\Lambda} (x), \hat{\Gamma} (X) = \hat{\Gamma} (x) \right\}. \quad (S6) \]

On this conditioning set, we have that \( S_{12} (X) = (S_{11} (x))^{\frac{1}{2}} \hat{\Lambda} (x) \hat{\Lambda} (x) \left( \hat{\Gamma} (x) \right)^T (S_{22} (x))^{\frac{1}{2}}. \)

Recalling the definition in (15), we see that \( S_{12} (X) = S_{12}' (x, R, \left\{ \hat{\lambda}_i (X) \right\}_{i=1}^{p_1}). \) Thus, on the conditioning set (S6), we have that

\[ S (X) = \begin{pmatrix} S_{11} (x) \\ S_{12}' (x, R, \left\{ \hat{\lambda}_i (X) \right\}_{i=1}^{p_1}) \end{pmatrix}^T \begin{pmatrix} S_{12}' (x, R, \left\{ \hat{\lambda}_i (X) \right\}_{i=1}^{p_1}) & S_{22} (x) \end{pmatrix} = S' (x, R, \left\{ \hat{\lambda}_i (X) \right\}_{i=1}^{p_1}), \]

where \( S' (x, R, \left\{ \hat{\lambda}_i (X) \right\}_{i=1}^{p_1}) \) was defined in Theorem 1. Therefore, we can rewrite \( p (x; R) \)
in (12) as

\[
p(x; R) = P_{H_0} \left( \prod_{i=1}^{p_1} \left( 1 - \left( \hat{\lambda}_i(X) \right)^2 \right) \right) \leq \prod_{i=1}^{p_1} \left( 1 - \left( \hat{\lambda}_i(x) \right)^2 \right) \quad \left( R \in \mathcal{P} \left( S'(x, R, \{\hat{\lambda}_i(X)\}_{i=1}^{p_1}) \right) \right),
\]

\[
S_{11}(X) = S_{11}(x), S_{22}(X) = S_{22}(x),
\]

\[
\hat{A}(X) = \hat{A}(x), \hat{\Gamma}(X) = \hat{\Gamma}(x).
\]

From Lemma 1 we have that under \( H_0 \), \( \{\hat{\lambda}_i(X)\}_{i=1}^{p_1} \) is independent of \( (S_{11}(X), S_{22}(X), \hat{A}(X), \hat{\Gamma}(X)) \). It follows that

\[
p(x; R) = P_{H_0} \left( \prod_{i=1}^{p_1} \left( 1 - \left( \hat{\lambda}_i(X) \right)^2 \right) \right) \leq \prod_{i=1}^{p_1} \left( 1 - \left( \hat{\lambda}_i(x) \right)^2 \right) \quad \left( R \in \mathcal{P} \left( S'(x, R, \{\hat{\lambda}_i(X)\}_{i=1}^{p_1}) \right) \right).
\]

Recall from Proposition 1 that under \( H_0 \), \( (\hat{\lambda}_1(X), \hat{\lambda}_2(X), \ldots, \hat{\lambda}_{p_1}(X)) \) has the density given in (7). Therefore, for \( (\lambda_1, \lambda_2, \ldots, \lambda_{r(P)}) \) with density given in (7), the first statement in Theorem 1 is established.

The proof of (16) is similar to the proof of Equation 11 in Theorem 1 of Gao et al. (2022). Using the probability integral transform we have that \( p(x; R) \) in (12) controls the selective type I error, i.e.

\[
P_{H_0} \left( p(x; R) \leq \alpha \left| R \in \mathcal{P}(S(X)), S_{11}(X) = S_{11}(x), S_{22}(X) = S_{22}(x), \hat{A}(X) = \hat{A}(x), \hat{\Gamma}(X) = \hat{\Gamma}(x) \right) \right) \leq \alpha.
\]

By the law of iterated expectation,

\[
P_{H_0} \left( p(x; R) \leq \alpha \left| R \in \mathcal{P}(S(X)) \right) \right) = \mathbb{E}_{H_0} \left( 1 \{ p(x; R) \leq \alpha \} \left| R \in \mathcal{P}(S(X)) \right) \right)
\]

\[
= \mathbb{E}_{H_0} \left( \mathbb{E}_{H_0} \left[ 1 \{ p(x; R) \leq \alpha \} \left| R \in \mathcal{P}(S(X)) \right), S_{11}(X) = S_{11}(x), S_{22}(X) = S_{22}(x), \hat{A}(X) = \hat{A}(x), \hat{\Gamma}(X) = \hat{\Gamma}(x) \right| R \in \mathcal{P}(S(X)) \right) \right)
\]

\[
\leq \mathbb{E}_{H_0} \left( \alpha \left| R \in \mathcal{P}(S(X)) \right) \right) = \alpha.
\]

This proves (16) and completes the proof of Theorem 1.
S3.1 Proof of Lemma 1

Proof. In what follows, we omit the dependence of \( \hat{A}(X), S_{11}(X), S_{22}(X), \hat{A}(X), \hat{T}(X) \) on \( X \). Recall that (4) can be written as \( S_{12}^W = S_{11}^{-1} S_{12} S_{22}^{-1} \overset{\text{CMD}}{=} \hat{A} \hat{A} \hat{T}^\top \). We first prove a technical lemma that will help us in establishing Lemma 1 going forward.

Lemma 2. Under \( H_0 : \Lambda = 0, \hat{A} \) and \( (S_{11}, \hat{A}, \hat{T}) \) are conditionally independent given \( X_2 \), i.e. \( \Lambda \perp (S_{11}, \hat{A}, \hat{T}) \mid X_2 \). Furthermore, \( (\hat{A}, S_{11}, \hat{A}, \hat{T}) \perp X_2 \).

Proof. Let us define \( \hat{W} := n S_{12} S_{22}^{-1} S_{21} = G \hat{T}^\top, \hat{G} := (\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_{p_2}) \), and \( \hat{T} := n S_{11} - n S_{12} S_{22}^{-1} S_{21} \). In order to obtain the conditional joint density of \( G \) and \( T \) given \( X_2 \), we first prove the following holds under \( H_0 : \Lambda = 0 \):

1) \( \hat{g}_i \mid X_2 \sim i.i.d. N(0, \Sigma_{11}), i = 1, 2, \ldots, p_2 \).
2) \( \hat{T} \mid X_2 \sim \text{Wishart}_{p_1}(\Sigma_{11}, n - 1 - p_2) \).
3) \( \hat{G} \perp \hat{T} \mid X_2 \).

Let \( H_n \) denote the centering matrix of size \( n \), i.e. \( H_n = I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top \). By construction, \( H_n \) is symmetric and idempotent. We have

\[
\hat{W} = X_1^\top H_n X_2 (X_2^\top H_n X_2)^{-1} X_2^\top H_n X_1 := X_1^\top Q_1 X_1, \quad Q_1 := H_n X_2 (X_2^\top H_n X_2)^{-1} X_2^\top H_n, \\
\hat{T} = X_1^\top H_n X_1 - X_1^\top H_n X_2 (X_2^\top H_n X_2)^{-1} X_2^\top H_n X_1 = X_1^\top (H_n - Q_1) X_1.
\]

We proceed as follows:

- Given \( X_2 \), under \( H_0 : \Lambda = 0 \), we have that \( X_1 \sim \mathcal{MN}_{n \times p_1}(0, I_n, \Sigma_{11}) \). We notice that \( Q_1 \) is a projection matrix onto the column space of \( H_n X_2 \), and therefore is idempotent with rank \( p_2 \). This implies that the eigenvalues of \( Q_1 \) are 1 with multiplicity \( p_2 \) and 0 with multiplicity \( n - p_2 \). Following the proof of Cochran’s Theorem in Theorem 3.4.4 (a) of Mardia et al. (1979), we have that \( X_1^\top Q_1 X_1 = \sum_{i=1}^{p_2} \hat{g}_i \hat{g}_i^\top \), with \( \hat{g}_i = X_1^\top v(i) \) where \( v(i) \) is the \( i \)th eigenvector of \( Q_1 \). Using Theorem 3.3.2 and 3.3.3 of Mardia et al. (1979), we have that \( \hat{g}_i \mid X_2 \sim i.i.d. N(0, \Sigma_{11}), i = 1, 2, \ldots, p_2 \). This proves 1).

- Next, we observe that \( H_n - Q_1 \) is an idempotent matrix, with rank \( (H_n - Q_1) = \text{trace}(H_n) - \text{trace}(Q_1) = n - 1 - p_2 \). Theorem 3.4.4 (b) of Mardia et al. (1979) implies \( T \mid X_2 \sim \text{Wishart}_{p_1}(\Sigma_{11}, n - 1 - p_2) \). This proves 2).

- We notice that \( (H_n - Q_1) Q_1 = H_n Q_1 - Q_1 = 0 \), where the first and the last equality follows from the idempotence of \( Q_1 \) and \( H_n \), respectively. Using Craig’s Theorem in Theorem 3.4.5 of Mardia et al. (1979), this implies \( \hat{W} \perp \hat{T} \mid X_2 \). As shown in the proof of 1), given \( X_2 \), \( G \) is a function of \( \hat{W} \). Hence, \( G \perp \hat{T} \mid X_2 \). This proves 3).
Combining 1)-3) above, we obtain the conditional joint density of $(\tilde{G}, \tilde{T})$, given $X_2$:

$$f \left( \tilde{G} = (\tilde{g}_1, \tilde{g}_2, \ldots, \tilde{g}_{p_2}), \tilde{T} | X_2 \right) \propto |\tilde{T}|^{\frac{n-p-1}{2}} \exp \left[ -\frac{1}{2} \left( \text{tr} \left( \Sigma_{11}^{-1} \tilde{T} \right) + \sum_{i=1}^{p_2} \tilde{g}_i \Sigma_{11}^{-1} \tilde{g}_i^\top \right) \right]$$

$$= |\tilde{T}|^{\frac{n-p-1}{2}} \exp \left[ -\frac{1}{2} \left( \text{tr} \left( \Sigma_{11}^{-1} \tilde{T} \right) + \sum_{i=1}^{p_2} \text{tr} \left( \Sigma_{11}^{-1} \tilde{g}_i \tilde{g}_i^\top \right) \right) \right]$$

$$= |\tilde{T}|^{\frac{n-p-1}{2}} \exp \left[ -\frac{1}{2} \left( \text{tr} \left( \Sigma_{11}^{-1} (\tilde{T} + \tilde{G} \tilde{G}^\top) \right) \right) \right].$$

Next, motivated by Hsu (1939), we consider the change of variable $(\tilde{G}, \tilde{T}) \mapsto (S_{11}, \hat{A}, \hat{\Lambda}, \hat{\Gamma})$. Here, $nS_{11} = \tilde{T} + \tilde{G} \tilde{G}^\top$ and $\hat{A} \hat{\Lambda} \hat{\Gamma}^\top = S_{12}^W = n^{-\frac{1}{2}} S_{11}^{-\frac{1}{2}} \tilde{G}$. From (2)-(4) in Ben-Israel (1999), we have

$$f \left( S_{11}, \hat{A}, \hat{\Lambda}, \hat{\Gamma} | X_2 \right) \propto f \left( \tilde{G}, \tilde{T} | X_2 \right) \times$$

$$\left| J_{(\tilde{G}, \tilde{T}) \rightarrow (S_{11}, \hat{A}, \hat{\Lambda}, \hat{\Gamma})} \left( J_{(\tilde{G}, \tilde{T}) \rightarrow (S_{11}, \hat{A}, \hat{\Lambda}, \hat{\Gamma})} \right)^\top \right|^{-\frac{1}{2}},$$

where $J_{(\tilde{G}, \tilde{T}) \rightarrow (S_{11}, \hat{A}, \hat{\Lambda}, \hat{\Gamma})}$ is the Jacobian matrix of the transformation $(\tilde{G}, \tilde{T}) \mapsto (S_{11}, \hat{A}, \hat{\Lambda}, \hat{\Gamma})$.

Plugging in $\tilde{G} = n^{\frac{1}{2}} S_{11}^{-\frac{1}{2}} \hat{\Lambda} \hat{\Lambda} \hat{\Gamma}^\top$ and $\tilde{T} = nS_{11} - \tilde{G} \tilde{G}^\top$ into $f \left( \tilde{G}, \tilde{T} | X_2 \right)$ shows that it factors across the variables $S_{11}, \hat{A}, \hat{\Lambda}$ and $\hat{\Gamma}$:

$$f \left( \tilde{G}, \tilde{T} | X_2 \right) \propto nS_{11} - nS_{11}^{\frac{1}{2}} \hat{\Lambda} \hat{\Lambda} \hat{\Gamma}^\top \hat{\Lambda} \hat{\Lambda} \hat{\Gamma}^\top S_{11}^{\frac{1}{2}} \left| S_{11} \right|^{\frac{n-p-1}{2}} \exp \left[ -\frac{1}{2} \left( \text{tr} \left( n \Sigma_{11}^{-1} S_{11} \right) \right) \right]$$

$$\propto \left| I_{p_1} - \hat{\Lambda} \hat{\Lambda}^\top \right|^{\frac{n-p-1}{2}} \left| S_{11} \right|^{-\frac{n-p-1}{2}} \exp \left[ -\frac{1}{2} \left( \text{tr} \left( n \Sigma_{11}^{-1} S_{11} \right) \right) \right]$$

$$= \left| I_{p_1} - \hat{\Lambda}^2 \right|^{\frac{n-p-1}{2}} \left| S_{11} \right|^{-\frac{n-p-1}{2}} \exp \left[ -\frac{1}{2} \left( \text{tr} \left( n \Sigma_{11}^{-1} S_{11} \right) \right) \right],$$

where the last equality follows by observing $\left| I - ZZ^\top \right| = \left| I - Z^\top Z \right|$, with the square matrix $Z = \hat{\Lambda} \hat{\Lambda}$. Thus, it remains to show that the Jacobian term in (S7) also factors. Now,

$$J_{(\tilde{G}, \tilde{T}) \rightarrow (S_{11}, \hat{A}, \hat{\Lambda}, \hat{\Gamma})} = \left( \frac{\partial S_{11}}{\partial \tilde{G}}, \frac{\partial S_{11}}{\partial \tilde{T}} \right),$$

where we write $\tilde{V} = (\hat{A}, \hat{\Lambda}, \hat{\Gamma})$. Noting that $\frac{\partial \tilde{V}}{\partial \tilde{T}} = 0$ and $\frac{\partial S_{11}}{\partial \tilde{G}} = n^{-1} I$,

$$J_{(\tilde{G}, \tilde{T}) \rightarrow (S_{11}, \hat{A}, \hat{\Lambda}, \hat{\Gamma})} \left( J_{(\tilde{G}, \tilde{T}) \rightarrow (S_{11}, \hat{A}, \hat{\Lambda}, \hat{\Gamma})} \right)^\top = \left( \frac{\partial S_{11}}{\partial \tilde{G}} \left( \frac{\partial S_{11}}{\partial \tilde{G}} \right)^\top + \frac{\partial \tilde{V}}{\partial \tilde{T}} \left( \frac{\partial \tilde{V}}{\partial \tilde{T}} \right)^\top \right) n^{-1} \frac{\partial S_{11}}{\partial \tilde{G}} n^{-2} I.$$
The determinant of a block matrix \( \begin{pmatrix} A & B \\ C & D \end{pmatrix} \), with an invertible \( D \), is given by \( |A - BD^{-1}C||D| \).

Applying this to \( J(G, \hat{T}) \Rightarrow (S_{11}, \hat{A}, \hat{A}, \hat{\Gamma}) \left( J(G, \hat{T}) \Rightarrow (S_{11}, \hat{A}, \hat{A}, \hat{\Gamma}) \right)^T \), we have

\[
\begin{vmatrix}
J(G, T) \Rightarrow (S_{11}, \hat{A}, \hat{A}, \hat{\Gamma}) \left( J(G, T) \Rightarrow (S_{11}, \hat{A}, \hat{A}, \hat{\Gamma}) \right)^T \\
\frac{\partial V}{\partial G} \left( \frac{\partial V}{\partial G} \right)^T
\end{vmatrix} = \begin{vmatrix}
\partial S_{12}^W / \partial G \partial S_{12}^W / \partial G \\
\partial S_{12}^W / \partial G \\
\partial S_{12}^W / \partial G
\end{vmatrix} = |S_{11}|^{-p_2} \left| \left( J(S_{11}^W) \Rightarrow (\hat{A}, \hat{A}, \hat{\Gamma}) \left( J(S_{11}^W) \Rightarrow (\hat{A}, \hat{A}, \hat{\Gamma}) \right)^T \right| |S_{11}|^{-p_2}
\]

(S9)

where we have used the chain rule for matrices, and properties of the determinant, and the fact that \( \frac{\partial V}{\partial S_{12}^W} = J(S_{11}^W) \Rightarrow (\hat{A}, \hat{A}, \hat{\Gamma}) \). In (10) of Rennie (2006), the Jacobian of the transformation \( (S_{11}^W) \Rightarrow (\hat{A}, \hat{A}, \hat{\Gamma}) \) was shown to be a separable function of \( (\hat{A}, \hat{A}, \hat{\Gamma}) \) not dependent on \( X_2 \). This implies that \( f(S_{11}, \hat{A}, \hat{A}, \hat{\Gamma}|X_2) \propto f(S_{11}|X_2) f(\hat{A}|X_2) f(\hat{\Gamma}|X_2) \), proving the conditional independence result. Lastly, observe that \( X_2 \) does not appear anywhere in the conditional density, and thus \( (S_{11}, \hat{A}, \hat{A}, \hat{\Gamma}) \perp \perp X_2 \).

From Lemma 2, it follows that \( \hat{\Lambda} \perp (S_{11}, \hat{A}, \hat{\Gamma}) \) and \( \hat{\Lambda} \perp X_2 \), which completes the proof of Lemma 1.

\[\square\]

### S4 Proof of Proposition 2

First, we assume that the variables are unordered. For \( \hat{\mathcal{P}} \in \mathcal{P}(\mathbf{S}(\mathbf{x})) \), we want to show that

\[
\hat{\mathcal{P}} \in \mathcal{P}(\mathbf{S}'(\mathbf{x}, \hat{\mathcal{P}}, \{\lambda_i\}_{i=1}^{r(\hat{\mathcal{P}})})) \iff \max_{i' \in \mathcal{P}, j' \in \mathcal{P}^c} \left| \mathbf{R}_{i',j'} \left( \mathbf{S}'(\mathbf{x}, \hat{\mathcal{P}}, \{\lambda_i\}_{i=1}^{r(\hat{\mathcal{P}})}) \right) \right| \leq c.
\]

(S10)

First we prove \( \hat{\mathcal{P}} \in \mathcal{P}(\mathbf{S}'(\mathbf{x}, \hat{\mathcal{P}}, \{\lambda_i\}_{i=1}^{r(\hat{\mathcal{P}})})) \implies \max_{i' \in \mathcal{P}, j' \in \mathcal{P}^c} \left| \mathbf{R}_{i',j'} \left( \mathbf{S}'(\mathbf{x}, \hat{\mathcal{P}}, \{\lambda_i\}_{i=1}^{r(\hat{\mathcal{P}})}) \right) \right| \leq c \). Following Algorithm 1, in order to compute \( \mathcal{P}(\mathbf{S}'(\mathbf{x}, \hat{\mathcal{P}}, \{\lambda_i\}_{i=1}^{r(\hat{\mathcal{P}})})) \), we consider the adjacency matrix \( \mathbf{D} \left( \mathbf{S}'(\mathbf{x}, \hat{\mathcal{P}}, \{\lambda_i\}_{i=1}^{r(\hat{\mathcal{P}})})) \right) \), where for \( i', j' \in \{1, 2, \ldots, p\} \),

\[
\mathbf{D}_{i',j'} \left( \mathbf{S}'(\mathbf{x}, \hat{\mathcal{P}}, \{\lambda_i\}_{i=1}^{r(\hat{\mathcal{P}})})) \right) := \mathbb{1} \left\{ \left| \mathbf{R}_{i',j'} \left( \mathbf{S}'(\mathbf{x}, \hat{\mathcal{P}}, \{\lambda_i\}_{i=1}^{r(\hat{\mathcal{P}})}) \right) \right| > c \right\}.
\]

If \( \hat{\mathcal{P}} \in \mathcal{P}(\mathbf{S}'(\mathbf{x}, \hat{\mathcal{P}}, \{\lambda_i\}_{i=1}^{r(\hat{\mathcal{P}})})) \), then in the undirected graph corresponding to \( \mathbf{D} \left( \mathbf{S}'(\mathbf{x}, \hat{\mathcal{P}}, \{\lambda_i\}_{i=1}^{r(\hat{\mathcal{P}})})) \), the nodes in \( \hat{\mathcal{P}} \) are not connected with any of the nodes in \( \hat{\mathcal{P}}^c \).
Hence,
\[ \hat{P} \in \mathcal{P}\left(S' \left( x, \hat{P}, \{\lambda_i\}_{i=1}^{r(\hat{P})}\right) \right) \]
\[ \Rightarrow D_{i',j'} \left(S' \left( x, \hat{P}, \{\lambda_i\}_{i=1}^{r(\hat{P})}\right) \right) = 0, \text{ for all } i' \in \hat{P}; j' \in \hat{P}^c \]
\[ \Rightarrow \left|R_{i',j'} \left(S' \left( x, \hat{P}, \{\lambda_i\}_{i=1}^{r(\hat{P})}\right) \right) \right| \leq c, \text{ for all } i' \in \hat{P}; j' \in \hat{P}^c \]
\[ \Rightarrow \max_{i' \in \hat{P}, j' \in \hat{P}^c} \left|R_{i',j'} \left(S' \left( x, \hat{P}, \{\lambda_i\}_{i=1}^{r(\hat{P})}\right) \right) \right| \leq c. \]  
(S11)

This completes the proof that LHS \( \Rightarrow \) RHS in (S10).

Next, we prove \( \max_{i' \in \hat{P}, j' \in \hat{P}^c} \left|R_{i',j'} \left(S' \left( x, \hat{P}, \{\lambda_i\}_{i=1}^{r(\hat{P})}\right) \right) \right| \leq c \Rightarrow \hat{P} \in \mathcal{P} \left(S' \left( x, \hat{P}, \{\lambda_i\}_{i=1}^{r(\hat{P})}\right) \right), \) for \( i', j' \in \hat{P} \) or \( i', j' \in \hat{P}^c. \) (S12)

Moreover, by Algorithm 1, \( \hat{P} \in \mathcal{P} \left(S' \left( x, \hat{P}, \{\lambda_i\}_{i=1}^{r(\hat{P})}\right) \right) \) implies \( \max_{i' \in \hat{P}, j' \in \hat{P}^c} \left|R_{i',j'} \left(S' \left( x, \hat{P}, \{\lambda_i\}_{i=1}^{r(\hat{P})}\right) \right) \right| \leq c. \) Combining this with (S12) implies that \( D \left(S' \left( x, \hat{P}, \{\lambda_i\}_{i=1}^{r(\hat{P})}\right) \right) = D \left(S' \left( x, \hat{P}, \{\lambda_i\}_{i=1}^{r(\hat{P})}\right) \right). \) Recalling that \( \mathcal{P} \left(S' \left( x, \hat{P}, \{\lambda_i\}_{i=1}^{r(\hat{P})}\right) \right) \) depends only on the adjacency matrix \( D \left(S' \left( x, \hat{P}, \{\lambda_i\}_{i=1}^{r(\hat{P})}\right) \right) \) (see Algorithm 1), it follows that \( \hat{P} \in \mathcal{P} \left(S' \left( x, \hat{P}, \{\lambda_i\}_{i=1}^{r(\hat{P})}\right) \right). \) This completes the proof in the case of the unordered variables.

In the case of ordered variables, to show that \( \max_{i' \in \hat{P}, j' \in \hat{P}^c} \left|R_{i',j'} \left(S' \left( x, \hat{P}, \{\lambda_i\}_{i=1}^{r(\hat{P})}\right) \right) \right| \leq c \) implies \( \hat{P} \in \mathcal{P} \left(S' \left( x, \hat{P}, \{\lambda_i\}_{i=1}^{r(\hat{P})}\right) \right), \) we also need to show that \( \hat{P} \) consists only of consecutive variables. This follows immediately from the fact that \( \hat{P} \in \mathcal{P} \left(S' \left( x, \hat{P}, \{\lambda_i\}_{i=1}^{r(\hat{P})}\right) \right). \) This completes the proof in the case of ordered variables.

S5 Proof of Proposition 3

From (15), \( S' \left( x, \hat{P}, \{\lambda_i\}_{i=1}^{r(\hat{P})}\right) \) is linear in \( \left(\lambda_1, \lambda_2, \ldots, \lambda_{r(\hat{P})}\right). \) Furthermore,
\( H := \text{diag} \left(S' \left( x, \hat{P}, \{\lambda_i\}_{i=1}^{r(\hat{P})}\right) \right) \) is not a function of \( \left(\lambda_1, \lambda_2, \ldots, \lambda_{r(\hat{P})}\right), \) since only the off-diagonal blocks of \( S' \left( x, \hat{P}, \{\lambda_i\}_{i=1}^{r(\hat{P})}\right) \) are functions of \( \left(\lambda_1, \lambda_2, \ldots, \lambda_{r(\hat{P})}\right). \) Hence,
\( R \left(S' \left( x, \hat{P}, \{\lambda_i\}_{i=1}^{r(\hat{P})}\right) \right) = H^{-\frac{1}{2}} S' \left( x, \hat{P}, \{\lambda_i\}_{i=1}^{r(\hat{P})}\right) H^{-\frac{1}{2}} \) is linear in \( \left(\lambda_1, \lambda_2, \ldots, \lambda_{r(\hat{P})}\right). \)
Furthermore, \( \mathbf{R}_{\hat{\rho}, \hat{\rho}^c} \left( \mathbf{S}' \left( \mathbf{x}, \hat{\rho}, \{ \lambda_i \}_{i=1}^{r(\hat{\rho})} \right) \right) \) is a submatrix of \( \mathbf{R} \left( \mathbf{S}' \left( \mathbf{x}, \hat{\rho}, \{ \lambda_i \}_{i=1}^{r(\hat{\rho})} \right) \right) \), and hence is linear in \( (\lambda_1, \lambda_2, \ldots, \lambda_{r(\hat{\rho})}) \).

### S6 Proof of Theorem 2

The conditioning set \( \mathcal{G}(\mathbf{x}; \hat{\rho}) \) in (14) can be rewritten as in (17). Thus,

\[
\mathcal{G}(\mathbf{x}; \hat{\rho}) = \left\{ \left( \lambda_1, \lambda_2, \ldots, \lambda_{r(\hat{\rho})} \right) \in [0, 1]^{r(\hat{\rho})} : \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{r(\hat{\rho})}, \right. \\
\left. \begin{pmatrix}
\mathbf{R}_{\hat{\rho}, \hat{\rho}^c} \left( \mathbf{S}' \left( \mathbf{x}, \hat{\rho}, \{ \lambda_i \}_{i=1}^{r(\hat{\rho})} \right) \right) \\
-\mathbf{R}_{\hat{\rho}, \hat{\rho}^c} \left( \mathbf{S}' \left( \mathbf{x}, \hat{\rho}, \{ \lambda_i \}_{i=1}^{r(\hat{\rho})} \right) \right)
\end{pmatrix} \leq c \mathbf{1}_{\text{card}(\hat{\rho})} \mathbf{1}^\top\left( \text{card}(\hat{\rho}) \right) \right\}.
\]

The proof follows by combining the \( 2r(\hat{\rho}) \left( p - r(\hat{\rho}) \right) \) inequalities resulting from the max term in (17) with the \( r(\hat{\rho}) + 1 \) inequalities associated with upper and lower limits and ordering restrictions on \( (\lambda_1, \lambda_2, \ldots, \lambda_{r(\hat{\rho})}) \), i.e. \( \lambda_1 \leq 1, \lambda_l - \lambda_{l-1} \leq 0, \) for \( l = 2, \ldots, r(\hat{\rho}) \), and \( -\lambda_{r(\hat{\rho})} \leq 0 \).

### S7 Computation of \( p(\mathbf{x}; \hat{\rho}) \) in (19)

Here, we elaborate on the discussion in Section 4. The subsections here follow the organization of Section 4. The overall procedure can be found in Algorithm S1.

#### S7.1 Proof of Proposition 4

Without loss of generality, we assume \( \text{card}(\hat{\rho}) = p - 1 \). Next, we characterize the conditioning set in (14) for \( \text{card}(\hat{\rho}) = p - 1 \). From Theorem 1, we observe that the conditioning set is a function of a modified version of the sample covariance matrix with a perturbed off-diagonal block. Hence we primarily focus on the off-diagonal block of the covariance matrix. We denote the vectors \( \mathbf{s}_{\hat{\rho}, \hat{\rho}^c}^W(\mathbf{x}) \) and \( \mathbf{S}_{\hat{\rho}, \hat{\rho}^c}(\mathbf{x}) \) as \( \mathbf{s}_{\hat{\rho}, \hat{\rho}^c}^W(\mathbf{x}) \) and \( \mathbf{s}_{\hat{\rho}, \hat{\rho}^c}(\mathbf{x}) \), respectively. We denote the scalar \( \mathbf{s}_{\hat{\rho}, \hat{\rho}^c}(\mathbf{x}) \) as \( \mathbf{s}_{\hat{\rho}, \hat{\rho}^c}(\mathbf{x}) \). Using (4) we have

\[
\mathbf{s}_{\hat{\rho}, \hat{\rho}^c}^W(\mathbf{x}) = \left( \mathbf{S}_{\hat{\rho}, \hat{\rho}^c}(\mathbf{x}) \right)^{-\frac{1}{2}} \mathbf{s}_{\hat{\rho}, \hat{\rho}^c}(\mathbf{x}) \mathbf{S}^{VD} \mathbf{\hat{A}}^P(\mathbf{x}) \mathbf{\hat{A}}^P(\mathbf{x})^\top. \tag{S13}
\]

As \( \mathbf{\hat{A}}^P(\mathbf{x}) \mathbf{\hat{A}}^P(\mathbf{x}) \mathbf{\hat{G}}^P(\mathbf{x})^\top \) is the SVD decomposition of the vector \( \mathbf{s}_{\hat{\rho}, \hat{\rho}^c}^W(\mathbf{x}) \), the following holds true in (S13):
1. \( \hat{\mathbf{A}}^P (\mathbf{x}) \) is a scalar equal to \( \left\| s_{\hat{\mathbf{p}}, \hat{\mathbf{p}}^c}^W (\mathbf{x}) \right\|_2 \). We denote this as \( \hat{\lambda}_1^p (\mathbf{x}) \).

2. \( \hat{\Gamma}^P (\mathbf{x}) = 1 \).

3. \( \hat{\mathbf{A}}^P (\mathbf{x}) = \frac{1}{\left\| s_{\hat{\mathbf{p}}, \hat{\mathbf{p}}^c}^W (\mathbf{x}) \right\|_2} s_{\hat{\mathbf{p}}, \hat{\mathbf{p}}^c}^W (\mathbf{x}) \).

Substituting these in (S13), we have

\[
s_{\hat{\mathbf{p}}, \hat{\mathbf{p}}^c}^W (\mathbf{x}) = \hat{\lambda}_1^p (\mathbf{x}) \sqrt{s_{\hat{\mathbf{p}}, \hat{\mathbf{p}}^c} (\mathbf{x})} \left( \mathbf{S}_{\hat{\mathbf{p}}, \hat{\mathbf{p}}} (\mathbf{x}) \right)^{\frac{1}{2}} \frac{1}{\left\| s_{\hat{\mathbf{p}}, \hat{\mathbf{p}}^c}^W (\mathbf{x}) \right\|_2} s_{\hat{\mathbf{p}}, \hat{\mathbf{p}}^c}^W (\mathbf{x}). \tag{S14}
\]

Next for \( 1 \leq \lambda_1 \leq 0 \), we focus on the perturbed off-diagonal block of the modified covariance matrix \( \mathbf{S}' (\mathbf{x}, \hat{\mathbf{p}}, \{\lambda_1\}) \) in (14). For card \( \hat{\mathbf{p}}^c \) = 1, \( \mathbf{S}'_{\hat{\mathbf{p}}, \hat{\mathbf{p}}^c} \) in (15) can be rewritten as:

\[
\mathbf{S}'_{\hat{\mathbf{p}}, \hat{\mathbf{p}}^c} (\mathbf{x}, \hat{\mathbf{p}}, \{\lambda_1\}) = \left( \mathbf{S}_{\hat{\mathbf{p}}, \hat{\mathbf{p}}} (\mathbf{x}) \right)^{\frac{1}{2}} \left[ \mathbf{A}^P (\mathbf{x}) \lambda_1 \left( \hat{\Gamma}^P (\mathbf{x}) \right)^{\top} \right] \left( \mathbf{S}_{\hat{\mathbf{p}}, \hat{\mathbf{p}}^c}^W (\mathbf{x}) \right)^{\frac{1}{2}}
\]

\[
\lambda_1 \sqrt{s_{\hat{\mathbf{p}}, \hat{\mathbf{p}}^c} (\mathbf{x})} \left( \mathbf{S}_{\hat{\mathbf{p}}, \hat{\mathbf{p}}} (\mathbf{x}) \right)^{\frac{1}{2}} \frac{1}{\left\| s_{\hat{\mathbf{p}}, \hat{\mathbf{p}}^c}^W (\mathbf{x}) \right\|_2} s_{\hat{\mathbf{p}}, \hat{\mathbf{p}}^c}^W (\mathbf{x}) \tag{S15}
\]

where the last equality follows from (S14). Next, we look at the off-diagonal blocks of the correlation matrices of \( \mathbf{S}' (\mathbf{x}, \hat{\mathbf{p}}, \{\lambda_1\}) \) and \( \mathbf{S} (\mathbf{x}) \). We recall that \( \mathbf{R} (\mathbf{S}) \) denotes the correlation matrix corresponding to the covariance matrix \( \mathbf{S} \), as defined in Section 3.1. For \( i' \in \hat{\mathbf{p}} \) and \( j' \in \hat{\mathbf{p}}^c \),

\[
\mathbf{R}_{i', j'} \left( \mathbf{S}' (\mathbf{x}, \hat{\mathbf{p}}, \{\lambda_1\}) \right) = \frac{\mathbf{S}'_{i', j'} (\mathbf{x}, \hat{\mathbf{p}}, \{\lambda_1\})}{\sqrt{\mathbf{S}'_{i', i'} (\mathbf{x}, \hat{\mathbf{p}}, \{\lambda_1\}) \mathbf{S}'_{j', j'} (\mathbf{x}, \hat{\mathbf{p}}, \{\lambda_1\})}}
\]

\[
= \frac{\lambda_1}{\lambda_1^p (\mathbf{x})} \mathbf{R}_{i', j'} (\mathbf{S} (\mathbf{x})) \tag{S16}
\]

where the second inequality follows from combining (S15) and the fact that the diagonal elements of \( \mathbf{S}' (\mathbf{x}, \hat{\mathbf{p}}, \{\lambda_1\}) \) are equal to those of \( \mathbf{S} (\mathbf{x}) \). Next we focus on characterizing
the conditioning set defined in (17).

\[ G(x; \hat{P}) = \left\{ \lambda_1 \in [0, 1] : c \geq \max_{i' \in \hat{P}, j' \in \hat{P}^c} |R_{i', j'}(S'(x, \hat{P}, \{\lambda_1\}))| \right\} \]

\[ = \left\{ \lambda_1 \in [0, 1] : c \geq \frac{\lambda_1}{\hat{\lambda}^P(x)} \max_{i' \in \hat{P}, j' \in \hat{P}^c} |R_{i', j'}(S(x))| \right\} \]

\[ = \left\{ \lambda_1 \in [0, 1] : \lambda_1 \leq \frac{c\hat{\lambda}^P(x)}{\max_{i' \in \hat{P}, j' \in \hat{P}^c} |R_{i', j'}(S(x))|} \right\} \]  \hspace{1cm} (S17)

\[ = \left\{ \lambda_1 \geq 0 : \lambda_1^2 \leq g_u(S(x), \hat{P}, c) \right\}, \]

where \( g_u(S(x), \hat{P}, c) := \min \left\{ 1, \left( \frac{c\hat{\lambda}^P(x)}{\max_{i' \in \hat{P}, j' \in \hat{P}^c} |R_{i', j'}(S(x))|} \right)^2 \right\} \).

We can write the denominator of the first equality of (19) as

\[ \int_{\lambda_1 \in g(x; \hat{P})} f(\lambda_1) d\lambda_1 = \int_{\{\lambda_1 > 0\} \cap \{\lambda_1^2 \leq g_u\}} f(\lambda_1) d\lambda_1 \]

\[ = \int_{\{\lambda_1^2 \leq g_u\}} f(\lambda_1) d\lambda_1 \]

\[ = \int_0^{g_u} f_{\frac{p-1}{2}, \frac{(n-p)}{2}}(t) dt, \]

where \( f_{\alpha, \beta}(\cdot) \) is the probability density function of the Beta \((\alpha, \beta)\) distribution. Here, the first equality follows from (S17), the second equality follows from the fact that \( f(\lambda_1) = 0 \) if \( \lambda_1 < 0 \), and the third equality follows from a change of variable and (iii) of Proposition 1.

Next, we concentrate on the numerator in (19). For \( r(\hat{P}) = 1 \),

\[ \left\{ \prod_{i=1}^{r(\hat{P})} (1 - \lambda_i^2) \leq \prod_{i=1}^{r(\hat{P})} (1 - \hat{\lambda}_{i}^P(x)^2) \right\} \]

is equivalent to \( \lambda_1^2 \geq \hat{\lambda}_{i}^P(x)^2 \). Proceeding as in the case of the denominator, \( p(x; \hat{P}) \) in (19) can be rewritten as

\[ p(x; \hat{P}) = \frac{\int_{\min\{g_u, \hat{\lambda}_i^P(x)^2\}}^{g_u} f_{\frac{p-1}{2}, \frac{(n-p)}{2}}(t) dt}{\int_0^{g_u} f_{\frac{p-1}{2}, \frac{(n-p)}{2}}(t) dt}, \]

where \( f_{\alpha, \beta}(\cdot) \) is the probability density function of the Beta \((\alpha, \beta)\) distribution.

**S7.2 Numerical integration to approximate (19) when \( r(\hat{P}) \) is small**

For \( r(\hat{P}) > 1 \), we opt for numerical integration. Given \( L \) and \( g \) (defined in Theorem 2), we evaluate \( p(x; \hat{P}) \) in (19) through the following steps:
1. We compute the convex hull representation of the convex polytope given by \( \{ \lambda : \mathbf{L}\lambda \leq \mathbf{g} \} \). This is achieved with the function \texttt{scdd} in the R package \texttt{rcdd} (Geyer et al., 2021).

2. Next, we find the Delaunay triangulation of the convex hull. We use the function \texttt{delaunayn} in the R package \texttt{geometry} (Habel et al., 2019) to do this.

3. Let \( T_1, T_2, \ldots, T_N \) be the simplices of the triangulation. We integrate the functions \( f(\lambda_1, \lambda_2, \ldots, \lambda_r(\hat{P})) \mathbf{1} \left\{ \prod_{i=1}^r (1 - \lambda_i^2) \leq \prod_{i=1}^r \left( 1 - \hat{\lambda}_i (x)^2 \right) \right\} \) (corresponding to the numerator in (19)) and \( f(\lambda_1, \lambda_2, \ldots, \lambda_r(\hat{P})) \) (corresponding to the denominator in (19)) on the obtained simplices and take their corresponding sums. We use the function \texttt{adaptIntegrateSimplex} in the R package \texttt{SimplicialCubature} (Nolan et al., 2021) for this purpose.

4. Finally we compute \( p(x; \hat{P}) \) as the ratio of these two approximations,

\[
p(x; \hat{P}) = \frac{\sum_{t=1}^N \int_{T_t} f(\lambda_1, \ldots, \lambda_r(\hat{P})) \mathbf{1} \left\{ \prod_{i=1}^r (1 - \lambda_i^2) \leq \prod_{i=1}^r \left( 1 - \hat{\lambda}_i (x)^2 \right) \right\} d\lambda_1 \ldots d\lambda_r(\hat{P})}{\sum_{t=1}^N \int_{T_t} f(\lambda_1, \ldots, \lambda_r(\hat{P})) d\lambda_1 \ldots d\lambda_r(\hat{P})}.
\]

\[(\text{S18})\]

S7.3 Monte Carlo approximation of (19) when \( r(\hat{P}) \) is large

We use \( B = 1,000 \) for the Monte Carlo approximation in Section 4.3, unless \( \sum_{b=1}^{1000} \mathbf{1} \{ \mathbf{L}\lambda^{(b)} \leq \mathbf{g} \} < 100 \), in which case we use a larger value of \( B \).
Algorithm S1 Computation of $p\left(\mathbf{x}; \hat{\mathcal{P}}\right)$ in (13)

Input: Sample covariance matrix $\mathbf{S}(\mathbf{x})$; a group of variables $\hat{\mathcal{P}} \in \mathcal{P}(\mathbf{S}(\mathbf{x}))$, where $\mathcal{P}(\mathbf{S}(\mathbf{x}))$ is computed as described in Algorithm 1, using a threshold $c$; the sample size $n$; the number of Monte Carlo replicates, $B$.

1: procedure
2: Compute $\hat{\Lambda}^{\hat{\mathcal{P}}}(\mathbf{x})$ from (4) with diagonal elements $\{\hat{\lambda}_i^{\hat{\mathcal{P}}}(\mathbf{x})\}_{i=1}^{r(\hat{\mathcal{P}})}$;
3: if $r(\hat{\mathcal{P}}) = 1$ then
4: Compute $g_u$ as defined in Proposition 4;
5: Compute $p(\mathbf{x}; \hat{\mathcal{P}})$ from (20);
6: if $1 < r(\hat{\mathcal{P}}) \leq 5$ then
7: Compute $\mathbf{L}(\mathbf{S}(\mathbf{x}), \hat{\mathcal{P}})$ and $g(\hat{\mathcal{P}}, c)$ as defined in Theorem 2;
8: Compute the convex hull representation of the polytope $\{\lambda : \mathbf{L}\lambda \leq g\}$;
9: Compute Delaunay triangulation of the convex hull;
10: Identify the simplices of the triangulation, $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_N$;
11: for $t = 1, 2, \ldots, N$ do
12: Compute \( \int_{\mathcal{T}_t} f \left( \lambda_1, \ldots, \lambda_{r(\hat{\mathcal{P}})} \right) d\lambda_1 \ldots d\lambda_{r(\hat{\mathcal{P}})} \)
13: Compute $p(\mathbf{x}; \hat{\mathcal{P}})$ from (S18);
14: if $r(\hat{\mathcal{P}}) > 5$ then
15: Compute $\mathbf{L}(\mathbf{S}(\mathbf{x}), \hat{\mathcal{P}})$ and $g(\hat{\mathcal{P}}, c)$ as defined in Theorem 2;
16: for $b = 1, 2, \ldots, B$ do
17: Simulate $\lambda^{(b)}$ from (ii) of Proposition 1;
18: Approximate $p(\mathbf{x}; \hat{\mathcal{P}})$ with $\hat{p}_{MC}(\mathbf{x}; \hat{\mathcal{P}})$ from (22).

Furthermore, due to numerical instability, the numerical integration approach described in Section 4.2 on rare occasions may produce estimates of $p\left(\mathbf{x}; \hat{\mathcal{P}}\right)$ that are outside of $[0, 1]$. In those cases, we use the Monte Carlo approach from Section 4.3.

S8 Simulation setup

S8.1 Generation of $\Sigma$

We generate $\Sigma$ as $\Sigma = \Sigma_0 + w \mathbf{1}\mathbf{1}^\top$, where $\Sigma_0$ is a $p \times p$ positive definite matrix, randomly generated using the function genPositiveDefMat(·) with default options in the R package clusterGeneration (Qiu and Joe (2020)), and $w$ is a Unif[0, 1] random variable. This code first generates uniformly distributed eigenvalues on $[1, 10]$, and then uses columns of a randomly generated orthogonal matrix as eigenvectors to construct the covariance matrix $\Sigma_0$. Since this produces $\Sigma_0$ with small off-diagonal elements (i.e. low signal strength), we increase the signal strength by adding $w \mathbf{1}\mathbf{1}^\top$ to $\Sigma_0$. 32
S8.2 Choice of threshold $c$

For each $\Sigma$, the threshold $c$ is adaptively chosen based on $\Sigma$. First, let $\tilde{c}$ be the threshold that produces exactly two connected components when we apply the thresholding procedure described in Section 3.1 on the population correlation matrix $\text{diag} (\Sigma)^{-\frac{1}{2}} \Sigma \text{diag} (\Sigma)^{-\frac{1}{2}}$. To account for the deviation of the sample correlation matrix from the population correlation matrix, we set the threshold $c(\Sigma) := \max \left\{ \tilde{c} + \sqrt{\log(p)/n}, 1 - 10^{-8} \right\}$. 