T2-DAG: a powerful test for differentially expressed gene pathways via graph-informed structural equation modeling

Jin Jin and Yue Wang

1Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, 615 N Wolfe St, Baltimore, Maryland 21205, U.S.A. jjin31@jhu.edu

2School of Mathematical and Natural Sciences, Arizona State University, 4701 W Thunderbird Rd, Glendale, AZ 85306, U.S.A. yue.wang.stat@asu.edu

Abstract

Motivation: A major task in genetic studies is to identify genes related to human diseases and traits to understand functional characteristics of genetic mutations and enhance patient diagnosis. Compared to marginal analyses of individual genes, identification of gene pathways, i.e., a set of genes with known interactions that collectively contribute to specific biological functions, can provide more biologically meaningful results. Such gene pathway analysis can be formulated into a high-dimensional two-sample testing problem. Given the typically limited sample size of gene expression datasets, most existing two-sample tests tend to have compromised powers because they ignore or only inefficiently incorporate the auxiliary pathway information on gene interactions.

Results: We propose T2-DAG, a Hotelling’s $T^2$-type test for detecting differentially expressed gene pathways, which efficiently leverages the auxiliary pathway information on
gene interactions from existing pathway databases through a linear structural equation model. We further establish its asymptotic distribution under pertinent assumptions. Simulation studies under various scenarios show that T2-DAG outperforms several representative existing methods with well-controlled type-I error rates and substantially improved powers, even with incomplete or inaccurate pathway information or unadjusted confounding effects. We also illustrate the performance of T2-DAG in an application to detect differentially expressed KEGG pathways between different stages of lung cancer.

**Availability:** The R package T2DAG is available on Github at https://github.com/Jin93/T2DAG.

**Contact:** jjin31@jhu.edu; yue.wang.stat@asu.edu

**Supplementary information:** Supplementary data are available at Bioinformatics online.

1 **Introduction**

With the continuous development of high-throughput sequencing technologies, there is an increasing need for powerful statistical methods for large-scale quantitative comparison between populations/conditions. For example, in gene expression data analysis, a major task is to identify genes from a large gene list that express differentially between different conditions, such as different disease statuses. Various approaches have been proposed for this task, providing critical insights into the molecular basis of many complex human traits/diseases; see Robinson et al. (2010), Love et al. (2014) and the references therein.

Besides separately analyzing individual genes, recently there is a growing interest in detecting differentially expressed gene pathways. Here, a gene pathway refers to a set of correlated genes with known interactions collectively contributing to specific biological functions. Many biologically meaningful gene pathways have been identified via experiments or statistical modeling, along with identified gene interactions, such as activation/expression or inhibition/repression of one gene by another. Public databases containing such pathways include Kyoto Encyclopedia of Genes and Genomes (KEGG, Kanehisa and Goto, 2000; Kanehisa et al., 2019), Gene Ontol-
Fig. 1 shows two directed acyclic graphs (DAGs) illustrating two KEGG gene pathways, which will be revisited in our data analysis in Section 4. $a \rightarrow b$ and $a \leftarrow b$ both indicate activation from gene $a$ to gene $b$; $a \rightarrow \rightarrow b$ and $a \leftarrow b$ indicate expression and inhibition, respectively, from gene $a$ to gene $b$. A common and notable feature of the two pathways is the high sparsity of the edges in the corresponding DAGs; that is, only a small proportion of the gene pairs are connected. Such sparsity pattern is also present in most identified gene pathways (Wille et al., 2004; Li et al., 2012; Chun et al., 2015).

Figure 1: Directed acyclic graphs (DAGs) describing gene interactions in an AGE-RAGE signaling pathway in diabetic complications (left panel) and a pathway for central carbon metabolism in cancer (right panel) from the KEGG database.

There are two main reasons for investigating differentially expressed gene pathways over individual genes. First, marginal tests for individual genes tend to have low statistical power due to limited sample sizes of typical gene expression datasets. On the other hand, gene pathway tests can achieve enhanced power by aggregating concerted signals from multiple related genes in the same pathway. Second, marginal analyses of individual genes often ignore interactions among genes, which potentially lead to spurious conclusions inconsistent with the underlying
biological processes. Since gene interactions can provide critical insights into the molecular mechanism of the progression of traits/diseases, efficiently leveraging the auxiliary information on gene interactions can facilitate the detection of disease- and trait-specific genes and yield biologically more meaningful results.

To formalize the problem, let $\mathcal{G} = (\mathcal{V}, \mathcal{D})$ denote the pathway of interest, where $\mathcal{V} = \{1, \ldots, p\}$ and $\mathcal{D}$ denote the gene set and edge set between genes, respectively. We assume that $\mathcal{G}$ is a DAG; i.e., all edges in $\mathcal{D}$ are directed edges with no loops. A DAG $\mathcal{G}$ can be summarized by an adjacency matrix $A = (a_{jk}) \in \mathbb{R}^{p \times p}$, where $a_{jk} = 1$ if the edge $j \to k$ is present in $\mathcal{D}$, in which case we call gene $j$ a parent of gene $k$ and gene $k$ a child of gene $j$, and $a_{jk} = 0$ otherwise. Although some identified gene pathways cannot be described by DAGs, they tend to have a small number of loops and/or undirected edges (see Table S1 for a summary of the KEGG pathways included in our data analysis). Thus, for such gene pathways, analyses based on their largest directed acyclic subgraphs are still feasible and informative. Approaches to directly handling non-DAG pathways are feasible but will require techniques different from the ones in this study, which we will briefly discuss in Section 5.

Assume we have observed expression levels of the genes in $\mathcal{V}$ for individuals in two different conditions. Specifically, let $x^{(l)}_{ij}$ denote the expression level of gene $j \in \mathcal{V}$ for individual $i \in \{1, \ldots, n_l\}$ in group $l \in \{1, 2\}$. We assume that the data are appropriately transformed so that $x^{(l)}_i = (x^{(l)}_{i1}, \ldots, x^{(l)}_{ip})^T$, $i = 1, \ldots, n_l$, are independent and identically distributed (i.i.d.) random vectors from a multivariate Gaussian distribution with mean $\mu_l$ ($l = 1, 2$) and covariance $\Sigma$. Besides observed gene expression data, there is auxiliary information on the presence/absence of directed relationships among genes (contained in $\mathcal{D}$), such as activation/inhibition of one gene by another. Our goal is then to test for the inequality of the two population mean, i.e., $H_0 : \mu_1 = \mu_2$ versus $H_a : \mu_1 \neq \mu_2$, based on $x^{(l)}_i$s, while efficiently leveraging the auxiliary information on gene interactions.

Conventional hypothesis testing for the inequality of two mean vectors has been thoroughly
investigated. A classical approach is the Hotelling’s $T^2$ test (Hotelling, 1931) with the test statistic

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{x}^{(1)} - \bar{x}^{(2)})^\top \hat{\Sigma}^{-1} (\bar{x}^{(1)} - \bar{x}^{(2)}),$$

where $\bar{x}^{(l)} = \frac{1}{n_l} \sum_{i=1}^{n_l} x^{(l)}_i$ and $\hat{\Sigma} = (n_1 + n_2 - 1)^{-1} \sum_{l=1}^{2} \sum_{i=1}^{n_l} (x^{(l)}_i - \bar{x}^{(l)}) (x^{(l)}_i - \bar{x}^{(l)})^\top$. When the number of genes ($p$) is smaller than the total sample size ($n_1 + n_2$), $T^2$ is well-defined; under $H_0$, $\frac{T^2 (n_1 + n_2 - p - 1)}{(p (n_1 + n_2 - 2))}$ follows an $F$-distribution with parameters $p$ and $n_1 + n_2 - 1 - p$. The $T^2$ is a sum-of-squares type of statistic powerful against “dense alternatives”, where the signals of differential expression are scattered among a large number of genes. This can often be true for gene pathway analyses, as genes in the same pathway tend to have concerted signals of expression that together form certain biological functions.

While Hotelling’s $T^2$ is well known as the uniformly most powerful invariant (UMPI) test, it is not applicable when $p$ is larger than $(n_1 + n_2)$, referred to as the high-dimensional setting hereafter. This is simply because in the high-dimensional setting the sample covariance matrix $\hat{\Sigma}$ becomes singular, making $T^2$ ill-posed. Research has been conducted extensively on extending $T^2$ to the high-dimensional setting. For example, Bai and Saranadasa (1996) and Srivastava and Du (2008) directly modified the ill-posed $T^2$ by replacing $\hat{\Sigma}$ in (1) with the identity matrix $I$ and the diagonal of $\hat{\Sigma}$, respectively. Lopes et al. (2011) proposed to first project the high-dimensional data onto a randomly constructed low-dimensional space and then perform the Hotelling’s $T^2$ test. Chen et al. (2011) replaced $\hat{\Sigma}$ with $(\hat{\Sigma} + \lambda I)$ for some regularization parameter $\lambda > 0$ and proposed a nonparametric testing procedure. Recently, Li et al. (2020) extended the work of Chen et al. (2011) and proposed an Adaptive Regularized Hotelling’s $T^2$ (ARHT) test that combines the test statistics corresponding to a set of $\lambda$s optimally chosen by a data-adaptive selection mechanism. Methods that are not direct modifications of Hotelling’s $T^2$ are also extensive. For example, Cai et al. (2014) proposed the CLX test with a supremum-type test statistic powerful against “sparse” alternatives. Gregory et al. (2015) proposed the GCT test that avoids the estima-
tion of $\hat{\Sigma}^{-1}$ by assuming an ordering of the variables so that the dependence between variables can be modeled according to their displacement. Xu et al. (2016) proposed the aSPU test, a robust adaptive test powerful against a wide range of alternatives.

While these existing methods can, to varying degrees, handle potentially high-dimensional gene expression data, they cannot incorporate the auxiliary information on gene interactions available from pathway databases. Incorporating such auxiliary information could substantially improve the test power, especially considering the inadequate sample size of a typical gene expression dataset. Previous efforts to address this issue include the NetGSA test (Shojaie and Michailidis, 2009 and Shojaie and Michailidis, 2010), which assumes that the magnitude of gene interactions is priorly known and incorporates such information using a linear mixed model. This approach has limited applicability in real applications because the magnitude of gene interactions is often unknown or only partially known and needs to be estimated using external data. Another approach proposed by Jacob et al. (2012), referred to as the Graph T2 test hereafter, projects data onto a low-dimensional space spanned by top eigenvectors of the graph Laplacian of the gene pathway and then performs the dimension reduction test in Lopes et al. (2011). Unlike the NetGSA test, Graph T2 can simply incorporate information on the presence/absence of gene interactions, but it may have compromised power if the signal is not coherent with the selected eigenvectors of the corresponding graph Laplacian.

To address these limitations of the existing methods, we propose T2-DAG, a novel test for identifying differentially expressed gene pathways, which efficiently leverages auxiliary information on gene interactions. Unlike the NetGSA test that requires additional information on the magnitude of gene interactions, the T2-DAG test only requires information on the presence/absence of gene interactions, which is more easily accessible in practice. Although the Graph T2 test also requires only the information on the presence/absence of gene interactions, it simply incorporates a few eigenvectors of the corresponding graph Laplacian. On the contrary, T2-DAG directly incorporates all gene interactions into the analysis with minimal information.
loss. Our main idea is to link the auxiliary information on gene interactions with (partial) correlations among genes through a linear structural equation model (SEM). Based on the SEM, we construct a novel pathway-informed estimator of the precision matrix of the genes via the LDL decomposition (Krishnamoorthy and Menon, 2013), where the ordering of the genes is naturally determined by the topological ordering of the pathway. We then construct a test statistic by replacing the ill-posed \( \hat{\Sigma}^{-1} \) in (1) with the novel estimator of the precision matrix followed by a \( Z \)-transformation, which is denoted by T2-DAG \( Z \). Under mild conditions, we prove that the proposed test statistic converges in distribution to the standard normal distribution. An alternative T2-DAG \( \chi^2 \) test can be constructed without the \( Z \)-transformation, which converges in distribution to \( \chi^2_{p} \) under more stringent conditions. We also derive the asymptotic power of the tests under local alternatives that characterize the true signals of differential expression in the gene pathway between two populations/conditions. Under various simulated data scenarios, the proposed T2-DAG tests show well-controlled type I error rates and substantially higher powers compared to multiple existing methods, even with partially missing or incorrect information on gene interactions, unadjusted confounding effects, or non-Gaussian error terms. Although theoretically limited, the T2-DAG \( \chi^2 \) test compares favorably to the T2-DAG \( Z \) test in the simulated scenarios, where it has slightly higher but still well-controlled type I error rates and higher powers. We also demonstrate the efficiency of the T2-DAG tests in a lung cancer-related application to detect differentially expressed gene pathways between different cancer stages.

2 Methods

Recall that our goal is to test for inequality of the mean expression levels of a gene pathway in two populations, \( H_0 : \mu_1 = \mu_2 \) versus \( H_a : \mu_1 \neq \mu_2 \), based on observed gene expression data, \( x_i^{(l)} = (x_{i1}^{(l)}, \ldots, x_{ip}^{(l)})^\top, i = 1, \ldots, n_l, l = 1, 2 \), while leveraging available auxiliary information on gene interactions described by a DAG \( \mathcal{G} = (\mathcal{V}, \mathcal{D}) \). Throughout the article, we use \( \xi_1 \perp \perp \xi_2 \) to
denote the independence between two random variables $\xi_1$ and $\xi_2$. We write $g(n) = O(f(n))$ if and only if $\lim_{n \to \infty} g(n)/f(n) < \infty$, and use $g(n) = o(f(n))$ to denote $\lim_{n \to \infty} g(n)/f(n) = 0$. We also write $f(n) \asymp g(n)$ if and only if $\lim_{n \to \infty} |f(n)/g(n)| = C$ for some constant $C > 0$. Let $I(\mathcal{A})$ denote the indicator function of the event $\mathcal{A}$; i.e., $I(\mathcal{A}) = 1$ if $\mathcal{A}$ is true, and $I(\mathcal{A}) = 0$ otherwise. We use lowercase letters, bold lowercase letters, and uppercase letters to denote scalars, vectors, and matrices, respectively.

We first define a new random variable, $z_i = (z_{i1}, \ldots, z_{ip})^\top$, based on the original gene expression data, $x_{il_i}, i = 1, \ldots, n_i, l = 1, 2$:

\[
z_i = \begin{cases} 
    x_{i1}^{(1)} - \mu_1 & i = 1, 2, \ldots, n_1 \\
    x_{i-n_1}^{(2)} - \mu_2 & i = n_1 + 1, \ldots, n_1 + n_2.
\end{cases}
\]

Given that $x_{il_i}$ are i.i.d. random vectors that follow a multivariate normal distribution $\mathcal{N}(\mu_l, \Sigma)$, we can easily check that $z_i \sim \mathcal{N}(0, \Sigma), i = 1, 2, \ldots, n_1 + n_2$. We propose to model each $z_i$ through the following SEM with a vector of Gaussian noises:

\[
z_i = Q^\top z_i + \epsilon_i, \tag{2}
\]

where $Q = (q_{jk}) \in \mathbb{R}^{p \times p}$ is an unknown coefficient matrix with $q_{jk}$ characterizing the magnitude of the effect of gene $j$ on gene $k$, and $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{ip})^\top$ are i.i.d. random vectors following $\mathcal{N}(0, R)$. We make the following assumption on $R$:

**Assumption 1.** $R = \text{diag}(r_1, \ldots, r_p)$.

Assumption 1 states that $\epsilon_{ij_1} \perp \epsilon_{ij_2}$ for all $j_1 \neq j_2$ and $i = 1, \ldots, n_1 + n_2$. This amounts to the assumption that there is no latent confounders, which is common in the existing SEM literature; see Loh and Bühlmann (2014), Peters and Bühlmann (2014), and the references therein. We next link the coefficient matrix $Q$ in model (2) with the DAG $\mathcal{G}$ through the following assumption.
Assumption 2. The support of \( Q \), i.e., the indices of nonzero entries in \( Q \), is informed by \( G \) through the corresponding adjacency matrix \( A = (a_{jk}) \), such that \( q_{jk} \neq 0 \) if and only if \( a_{jk} \neq 0 \).

Assumption 2 is only required for theoretical analysis. Simulation studies in Section 3 will allow the auxiliary DAG to have missing or mis-specified edges. Since \( G \) is a DAG, we can label the variables in \( \mathcal{V} \) according to the topological ordering of \( G \) so that \( Q \) is a strictly upper triangular matrix, i.e., an upper triangular matrix with the diagonal entries being 0 (Loh and Bühlmann, 2014). As a result, model (2) can be rewritten as the following \( p \) linear models:

\[
  z_{ij} = f_j(\epsilon_{i1}, \ldots, \epsilon_{ij}), \quad j = 1, \ldots, p, \quad i = 1, \ldots, n_1 + n_2,
\]

where \( \{f_j(\cdot)\}_{j=1}^p \) are linear functions that are determined by \( Q \). Thus, for \( 1 \leq j_2 < j_1 \leq p \) and each \( i \), one can check that \( \text{Cov}(\epsilon_{ij_1}, z_{ij_2}) = 0 \), indicating that \( \epsilon_{ij_1} \perp z_{ij_2} \) under Gaussianity. This property will play an important role in establishing the asymptotic null distribution of our test statistic.

We next construct DAG-informed estimators of \( \Sigma \) and \( \Sigma^{-1} \). From model (2), we can obtain

\[
  \epsilon_i = (I - Q^T)z_i, \quad i = 1, \ldots, n_1 + n_2.
\]

Applying the covariance function to both sides of this equation yields

\[
  R = (I - Q^T)\Sigma(I - Q).
\]

Since \( Q \) is strictly upper triangular, we know that \( I - Q \) is invertible, leading to

\[
  \Sigma = (I - Q^T)^{-1}R(I - Q)^{-1}.
\]

Taking the inverse of both sides of (3) gives

\[
  \Sigma^{-1} = (I - Q)R^{-1}(I - Q)^T.
\]

Equations (3) and (4) serve as the basis for the construction of the DAG-informed estimators of \( \Sigma \) and \( \Sigma^{-1} \), which are critical to the proposed test statistic. Note that (3) and (4) are, respectively, LDL-type decompositions of \( \Sigma \) and \( \Sigma^{-1} \); similar decompositions have been employed in the
literature for estimating high-dimensional covariance/precision matrices (Wu and Pourahmadi, 2003; Huang et al., 2006). The rationale behind the LDL decomposition is that there exists some natural ordering of the variables, which, in the case of gene pathways, is determined by the topological ordering of the corresponding graph $G$.

We observe from equations (3) and (4) that $\Sigma$ and $\Sigma^{-1}$ are functions of $Q$ and $R$. We then construct estimators for $Q$ and $R$. Letting $q_j$ denote the $j$-th column of $Q$, we decompose model (2) into $p$ sub-models,

$$z_{ij} = z_i^T q_j + \epsilon_{ij}, \ j = 1, \ldots, p.$$  \hspace{1cm} (5)

To estimate the unknown parameters in (5), it is natural to consider the following square loss function for each sub-model $j$:

$$\sum_{l=1}^{2} \sum_{i=1}^{n_l} \left\{ x^{(l)}_{ij} - \mu_{lj} - \left( x^{(l)}_i - \mu_l \right)^T q_j \right\}^2 = \sum_{l=1}^{2} \sum_{i=1}^{n_l} \left\{ x^{(l)}_{ij} - \gamma_{lj} - x^{(l)}_i q_j \right\}^2,$$  \hspace{1cm} (6)

where $\gamma_{lj} = \mu_{lj} + \mu_l^T q_j$. Letting $S_j = \{ k : a_{kj} = 1 \}$, one can see that under Assumption 2, $q_j$ has $|S_j|$ nonzero entries with $|S_j| \leq j - 1$ for $j > 1$ and that $q_1 = 0$. Since real gene pathways are far less dense than fully connected networks (Wille et al., 2004; Li et al., 2012; Chun et al., 2015), we make the following assumption on $S_j$:

**Assumption 3.** $d = o(n_1 + n_2)$ as $n_1, n_2 \rightarrow \infty$, where $d = \max_j |S_j|$.

This assumption appears to be reasonable for all 206 pathways considered in our data analysis in Section 4 (see Table S1 for details). Under Assumptions 1-3, we can estimate $\gamma_{lj}$ and the nonzeros entries of $q_j$ in (6) using the ordinary least squares (OLS). Specifically, denote by $\mathbf{1}_{n_1+n_2}$ the $(n_1 + n_2)$-dimensional vector of all ones and let $g = (g_1, \ldots, g_{n_1+n_2})^T$ denote the group indicator; that is, $g_i$ equals 1 for $i = 1, \ldots, n_1$ and equals 0 for $i = n_1 + 1, \ldots, n_1 + n_2$. 


Then, letting $X = \begin{bmatrix} x_1^{(1)} & \cdots & x_{n_1}^{(1)} & x_1^{(2)} & \cdots & x_{n_2}^{(2)} \end{bmatrix}^\top$, we rewrite (6) in the following matrix form

$$
\| X_j - \theta_{1j} \mathbf{1}_{n_1+n_2} - \theta_{2j} \mathbf{g} - X_{S_j} \hat{q}_{j,S_j} \|_2^2,
$$

where $X_j$ denotes the $j$-th column of $X$, $X_{S_j}$ denotes the sub-matrix of $X$ with columns indexed by $S_j$, and $q_{j,S_j}$ denotes the sub-vector of $q_j$ with entries indexed by $S_j$ that are non-zero according to the DAG and need to be estimated. Comparing (7) with (6), one can see that $\gamma_{1j} = \theta_{1j} + \theta_{2j}$ and $\gamma_{2j} = \theta_{1j}$. Thus, letting $W = [\mathbf{1}_{n_1+n_2} \mathbf{g}]$, we can obtain the OLS estimator of $q_{j,S_j}$ and $(\theta_{1j}, \theta_{2j})^\top$:

$$
\hat{q}_{j,S_j} = \left( X_{S_j}^\top (I_{n_1+n_2} - \mathcal{P}_W) X_{S_j} \right)^{-1} X_{S_j}^\top (I_{n_1+n_2} - \mathcal{P}_W) X_j,
$$

$$
(\hat{\theta}_{1j}, \hat{\theta}_{2j})^\top = (W^\top W)^{-1} W^\top (X_j - X_{S_j} \hat{q}_{j,S_j}),
$$

where $\mathcal{P}_W = W (W^\top W)^{-1} W^\top$ is the orthogonal projection matrix onto the column space of $W$.

For ease of notation, we also denote the OLS estimator of $q_{kj}$ by $\hat{q}_{kj}$ which equals 0 if $k \notin S_j$, and define the estimator of $Q$ as $\hat{Q} = (\hat{q}_{jk})_{j,k=1,\ldots,p}$. Given $\hat{Q}$, we then construct a “plug-in” estimator of $R$ denoted by $\hat{R} = \text{diag}(\hat{r}_1, \ldots, \hat{r}_p)$, with

$$
\hat{r}_j = \frac{1}{n_1 + n_2 - |S_j| - 4} \left\| X_j - \hat{\theta}_{1j} \mathbf{1}_{n_1+n_2} - \hat{\theta}_{2j} \mathbf{g} - X_{S_j} \hat{q}_{j,S_j} \right\|_2^2.
$$

The denominator, $(n_1 + n_2 - |S_j| - 4)$, makes $\hat{r}_j$ slightly different from the classical OLS estimator of $r_j$. Although it may seem non-intuitive, we show in Lemma 1.3 in the supplementary material that $\mathbb{E}[\hat{r}_j^{-1}] = r_j^{-1}$; this property is critical for establishing the asymptotic null distribution of our test statistic.

By replacing $Q$ and $R$ in (3) and (4) with $\hat{Q}$ and $\hat{R}$, we have the following DAG-informed
empirical estimators of $\Sigma$ and $\Sigma^{-1}$:

$$\hat{\Sigma}_{\text{DAG}} = (I - \hat{Q}^\top)^{-1}\hat{R}(I - \hat{Q})^{-1}$$

and

$$\hat{\Sigma}_{\text{DAG}}^{-1} = (I - \hat{Q})\hat{R}^{-1}(I - \hat{Q})^\top.$$  

Finally, we propose our DAG-informed $Z$ test statistic:

$$T_{\text{DAG}-Z} = \left(2^{p-1/2}\left(T_{\text{DAG}-\chi^2}^2 - p\right)\right),$$  

where $T_{\text{DAG}-\chi^2}^2$ is a DAG-informed Hotelling’s $T^2$-type test statistic,

$$T_{\text{DAG}-\chi^2}^2 = N(\bar{x}^{(1)} - \bar{x}^{(2)})^\top\hat{\Sigma}_{\text{DAG}}^{-1}(\bar{x}^{(1)} - \bar{x}^{(2)}),$$  

with $N = n_1n_2/(n_1+n_2)$, which is an alternative test statistic for testing $H_0$ that will be discussed in Remark 2. As such, we call the proposed testing procedure T2-DAG hereafter.

Let $N_e = \sum_{j=1}^p |S_j|$ and $p_0 = \sum_{j=1}^p I(|S_j| > 0)$ denote the number of edges in $D$ and the number of genes with at least one parent gene, respectively. The following result characterizes the asymptotic null distribution of $T_{\text{DAG}-Z}$.

**Theorem 1.** Suppose Assumptions 1-3 hold. As $n_1, n_2 \to \infty$, if $d \log p_0 = O(n_1 + n_2)$ and $N_e = o\left(\sqrt{p_1(n_1 + n_2)}\right)$, then under $H_0$, we have

$$T_{\text{DAG}-Z} \to \mathcal{N}(0, 1) \quad \text{in distribution, as } n_1, n_2 \to \infty.$$  

(11)

Theorem 1 directly yields an asymptotically valid two-sided p-value for $T_{\text{DAG}-Z}$ for testing $H_0 : \mu_1 = \mu_2$, that is, $p_{\text{DAG-Z}} = \mathbb{P}(|z| > |T_{\text{DAG-Z}}|)$, where $z$ is a random variable following the standard normal distribution. The assumptions in Theorem 1 often hold for typical gene pathways. In particular, $N_e = o\left(\sqrt{p_1(n_1 + n_2)}\right)$ implies that the total number of edges in the gene pathway is small compared to the square root of the number of genes multiplied by the total
sample size, which is a weak assumption in real applications. See Table S1 for information on the real pathways considered in our data analysis.

To examine the asymptotic power of the proposed test, we first define the following two distance measures between the mean vectors of the two groups:

1. Kullback-Leibler divergence: \( D_{KL}(\mu_1, \mu_2) = (\mu_1 - \mu_2)^T (I - Q) R^{-1} (I - Q)^T (\mu_1 - \mu_2) \).

2. Pathway-specific divergence: \( D_{DAG}(\mu_1, \mu_2) = \sum_{j: |S_j| > 0} (\mu_1 - \mu_2)^T S_j \left\{ \Sigma_{S_j, S_j} \right\}^{-1} (\mu_1 - \mu_2) S_j \), where \( (\mu_1 - \mu_2)_{S_j} \) denotes the sub-vector of \( \mu_1 - \mu_2 \) with entries indexed by \( S_j \), and \( \Sigma_{S_j, S_j} \) denotes the sub-matrix of \( \Sigma \) with rows and columns both indexed by \( S_j \).

The following result characterizes the asymptotic power of the proposed test under local alternatives regarding \( D_{KL}(\mu_1, \mu_2) \) and \( D_{DAG}(\mu_1, \mu_2) \).

**Theorem 2.** Suppose all the assumptions in Theorem 1 hold. If \( D_{DAG}(\mu_1, \mu_2) \asymp p^\beta \) and \( D_{KL}(\mu_1, \mu_2) \asymp p^{1/2} N^{-\gamma} \) for some \( 0 < \beta < 1/2 \) and \( \gamma > 1/2 \), then for all \( \alpha \in (0, 1) \),

\[
\lim_{n_1, n_2 \to \infty} \left[ \mathbb{P}(p_{DAG-Z} \leq \alpha) - \mathbb{P}(|Z| \leq -z_{\alpha/2} + \frac{ND_{KL}(\mu_1, \mu_2)}{\sqrt{2p}}) \right] \geq 0,
\]

where \( z \) is a standard normal random variable. In particular, if \( 1/2 < \gamma < 1 \), then we have \( \lim_{n_1, n_2 \to \infty} \mathbb{P}(p_{DAG-Z} \leq \alpha) = 1 \).

**Remark 1.** The proposed T2-DAG test is built upon the OLS estimators of \( Q \) and \( R \). When the OLS estimators are ill-posed (e.g., when Assumption 2 is violated), one can instead use penalized estimators such as the lasso estimator (Tibshirani, 1996) to estimate \( Q \) and \( R \), and the test statistic can be obtained by replacing \( \hat{Q} \) and \( \hat{R} \) in (10) with such penalized estimators. However, we may need alternative techniques to establish the asymptotic distribution of the resulting test statistic. This is because the current proof heavily depends on the theoretical property specific to the OLS estimator; that is, the OLS coefficient estimator is independent of the OLS variance estimator. Please refer to Lemma 1.3 in Section 1 of the supplementary material for more details.

**Remark 2.** The statistic \( T_{DAG-\chi^2}^2 \) in (10) defines an alternative test statistic for testing \( H_0 \). We
show in the proof of Theorem 1 (see Section 1 of the supplementary material) that $T_{DAG-\chi^2}^2$ converges to the chi-squared distribution with $p$ degrees of freedom if

\[ p = o(n_1 + n_2), \quad N_e = o(n_1 + n_2), \quad \text{and} \quad d \log p_0 = O(n_1 + n_2). \]

(12)

Note that the conditions in (12) are more stringent than those required in Theorem 1 for $T_{DAG-Z}$; that is,

\[ N_e = o(\sqrt{p}(n_1 + n_2)) \quad \text{and} \quad d \log p_0 = O(n_1 + n_2). \]

(13)

Compared to (12), the conditions in (13) allow the number of genes in the pathway ($p$) to be asymptotically larger than the total sample size ($n_1 + n_2$), which is theoretically important in high-dimensional statistics. Moreover, the conditions in (13) allow the total number of edges in the pathway ($N_e$) to grow with the number of genes ($p$), which is a reasonable situation in real applications, while the conditions in (12) don’t.

Due to these theoretical limitations of $T_{DAG-\chi^2}^2$, we adopt the Z-transformed statistic $T_{DAG-Z}$ in this manuscript. However, as shown in Section 3, $T_{DAG-\chi^2}^2$ compares favorably to $T_{DAG-Z}$ in terms of the finite-sample performance under various simulated scenarios: it yields well-controlled type-I error rates and higher powers than $T_{DAG-Z}$. It would be interesting to further study the gap between the theoretical and numerical properties of the statistic $T_{DAG-\chi^2}^2$.

3 Simulation Studies

3.1 Set-up

We illustrate the performance of the proposed T2-DAG test with simulation studies under various settings of sample size ($n_1 = n_2$), dimension ($p$), number of children nodes, i.e., genes that have at least one parent gene ($p_0$), and the maximum number of directed edges toward one gene ($d$).
We applied both $T_{DAG-Z}$ in (9) and $T^2_{DAG-\chi^2}$ in ((10)) for the proposed T2-DAG test, and denote the corresponding tests by T2-DAG $Z$ and T2-DAG $\chi^2$, respectively. The data was simulated as follows. We generated two groups of observations according to $x^{(l)}_i = \mu_l + (I - Q^T)^{-1}e_i, l = 1, 2, i = 1, \ldots, n_l$. Specifically, we first characterized the support of $Q$ using a binary adjacency matrix $A = (a_{ij})_{p \times p}$ generated according to the characteristics (e.g., $p, p_0$, and $d$) of the real KEGG pathways included in our real data analysis (see Table S1 for details). We considered two scenarios regarding $|S_j| = \sum_{k=1}^p I(a_{kj} = 1)$ and $p_0 = \sum_{j=1}^p I(|S_j| > 0)$. In the first scenario, we set $p_0 = 0.6p$ and $|S_j| = 1 + \zeta_j$, where $\zeta_j \sim NB(3, 0.6)$. This resulted in a value of $d$ between 10 and 15. This reflects the scenario where there are condensed gene interactions towards a relatively small number of genes. In the second scenario, we set $p_0 = 0.8p$ and $\zeta_j \sim NB(3, 0.8)$ which resulted in a value of $d$ between 3 and 7. This is aligned with real scenarios where gene pathways with larger $p_0$ tend to have smaller $d$, and it reflects the scenario of less condensed gene interactions where more genes in the gene set are affected by other genes (larger $p_0$) but each of them are affected by fewer genes (smaller $d$). Next, we generated an initial coefficient matrix, $Q^0$, with the $(i, j)$-th entry equal to $(-1)^{v_{ij}a_{ij}}$ where $v_{ij} \sim Bernoulli(0.5)$. We further rescaled $Q^0$ to obtain the final coefficient matrix, $Q = Q^0 / (\kappa \|Q^0\|_2)$ with $\kappa = 1.5$. We then generated $e_i$ independently from $N(0, R)$, $i = 1, \ldots, n_1 + n_2$, where $R = \text{diag}(r_0)$ with $r_0 = 0.2$. Without loss of generality, we set $\mu_1 = 0$ and $\mu_2 = (\mu_{2,1}, \ldots, \mu_{2,p})^T$ with $\mu_{2,j} = \delta I(j \leq q)$, where $\delta$ characterizes the signal strength, i.e., the magnitude of differential gene expression, and $q$ controls the proportion of genes in the pathway that express differentially between the two groups.

We compared the proposed T2-DAG test with several existing tests, including the Hotelling’s $T^2$ test (Hotelling, 1931), Graph T2 test (Jacob et al., 2012), BS (Bai and Saranadasa, 1996), Ch-Q (Chen and Qin, 2010), SK (Srivastava and Kubokawa, 2013), CLX (Cai et al., 2014), GCT (Gregory et al., 2015), aSPU (Xu et al., 2016), and ARHT (Li et al., 2020). The Hotelling’s
The $T^2$ test was only applied under the scenarios where $p < n_1 + n_2$. Besides information on the existence/absence of gene interactions, the Graph T2 test can also account for the signs of the interactions, such as activation (+) or inhibition (-). We therefore implemented Graph T2 based on top $k$ eigenvectors of the graph Laplacian matrix of the signed graph. We set the number of eigenvectors used to $k = 0.2p$ as suggested in Jacob et al. (2012). It has also been noted that Graph T2 is robust to the choice of $k$ (Lopes et al., 2011; Jacob et al., 2012). For GCT, we conducted the moderate-$p$ version of the test given the dimensions of the simulated datasets, and set the lag window size to $L = (2/3)p^{1/2}$ as suggested in Gregory et al. (2015).

For aSPU, the index of the sum-of-powers test, $\gamma$, was chosen adaptively from $\{1, 2, \ldots, 6, \infty\}$ (Xu et al., 2016). The ARHT test was implemented using a probabilistic prior model specified by the canonical weights, $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$, for $(I_p, \Sigma, \Sigma^2)$, with the type 1 error rate calibrated using cube-root transformation, which are the default settings of ARHT in the “ARHT” R package. All tests were conducted under a significance level of $\alpha = 0.05$.

### 3.2 Results given fully informative pathway information

We first implemented the T2-DAG test leveraging complete and correct edge information of the pathways summarized by the true adjacency matrix $A$. Fig. 2 shows the empirical rejection percentages (i.e., type I error rates when $\delta = 0$ and powers when $|\delta| > 0$) of all the tests when $n_1 = n_2 = 50$ and $p = 40$ or 100. Additional simulation results for $n_1 = n_2 = 50$ and $p = 20$, $n_1 = n_2 = 100$ and $p = 100$ or 300 have similar patterns and are summarized in Figs. S1-S3.

All tests yield well-controlled type I error rates, except for GCT which has substantial inflation in type I error rates in the lower-dimensional case ($n_1 = n_2 = 50$, $p = 40$). This is possibly because GCT assumes an ordering of the genes such that the dependency between genes can be modeled according to their displacement (Gregory et al., 2015), whereas in our simulated data we did not introduce this unique structure. This violation may lead to inflated type I error rates especially when $p$ is small.
Among the existing tests, CH-Q gives the best overall performance with a power higher than or equal to that of the Hotelling’s $T^2$ test (when $n_1 + n_2 > p$). The ARHT test shows a power that is similar to or slightly lower than that of the CH-Q test but higher than that of all other existing tests under all simulated scenarios, except for one scenario where aSPU has a slightly higher power (see Fig. 2, row 1, column 3). The Graph T2 test shows relatively low power in all scenarios. This is possibly because the simulated signal distribution is less coherent with the top eigenvectors of the graph Laplacian matrix. The power of the supremum-based CLX test is competitive to that of CH-Q and SK in sparse-signal scenarios (Fig. 2, columns 1 and 3), but is outperformed by most of the other tests under dense signals (Fig. 2, columns 2 and 4). The GCT test has an extremely low power in the sparse-signal scenarios (Fig. 2, columns 1 and 3), where the power is below 0.15 and even decreases as signal strength ($|\delta|$) increases. The proposed T2-DAG tests have substantially higher powers compared to the other tests in all simulated scenarios, except when signals are weak such that all tests have powers below 0.15. Compared to the T2-DAG $Z$ test that has a type I error rate strictly controlled at around $\alpha = 0.05$, T2-DAG $\chi^2$ has a
slightly higher but still well-controlled type I error rate and a higher power under all simulated scenarios. Furthermore, compared to the scenarios of $p_0 = 0.6p$ where there are condensed gene interactions towards a relatively small number of genes (Fig. 2, columns 1-2), in the scenarios of $p_0 = 0.8p$ where there are scattered gene interactions towards a larger number of genes (Fig. 2, columns 3-4), the two T2-DAG tests show more obvious advantages compared to the existing tests with strict type I error control and more improvement in power.

3.3 Results given partially incorrect edge information

We next discuss a more realistic scenario where the auxiliary edge information comes with errors, such as having missing and/or redundant edges. Recall that $N_e = \sum_{i=1}^{p} |S_j|$ denotes the total number of edges in the gene pathway. Besides the true adjacency matrix $A$, we consider two types of mis-specified auxiliary DAGs with the following adjacency matrices: (1) $A_1$ with a random set of $0.4N_e$ edges missing, and (2) $A_2$ with a random set of $0.4N_e$ redundant edges that do not exist in $A$. We implemented the proposed T2-DAG test based on $A$, $A_1$ or $A_2$, and implemented Graph T2 only based on the signed graph corresponding to the true adjacency matrix $A$.

Fig. 3 shows the empirical type I error rates and powers of all the tests given $n_1 = n_2 = 50$ and $p = 100$. Additional simulation results for $n_1 = n_2 = 50$ and $p = 50$, $n_1 = n_2 = 100$ and $p = 100$ or 300 are summarized in Figs. S4-S6. Although having incorporated the correct edge information as well as the signs of interactions, Graph T2 still has a much lower power compared to T2-DAG. Among the other existing tests that do not incorporate external graphical information, the ARHT test shows the highest power which is almost the same as that of CH-Q under relatively weak signals ($|\delta|$) and slightly higher than that of CH-Q under stronger signals. Similar to what we observed in Section 3.2, T2-DAG $\chi^2$ has well-controlled but slightly higher type I error rates and substantially higher power than T2-DAG $Z$. Using either of the two test statistics, the T2-DAG test based on pathway information with redundant edges (T2-DAG redundant) has slightly higher inflation in type I error rates, and the one based on partially missing
Figure 3: Empirical type I error rates (δ = 0) and powers (|δ| > 0) of T2-DAG Z (top panel) and T2-DAG χ² (bottom panel) under n₁ = n₂ = 50 and p = 100, with the number of children nodes set to p₀ = 0.8p and the number of non-zero signals set to q = 0.1p (column 1) or 0.5p (column 2). T2-DAG was applied based on A (“T2-DAG Correct”), A₁ (“T2-DAG Missing”), or A₂ (“T2-DAG Redundant”). Graph T2 was applied based on the true adjacency matrix A.

edge information (T2-DAG Missing) has slightly compromised powers. Nonetheless, the overall performance of T2-DAG is similar across the three scenarios. This indicates the robustness of the proposed T2-DAG test with respect to potential errors in the auxiliary pathway information.

3.4 Results in the presence of unadjusted confounding effects

So far, we have considered independent error terms ϵₖ : j = 1, . . . , p, with a diagonal covariance matrix R. However, when there exist latent confounders affecting gene interactions, the error terms may be correlated, leading to a non-diagonal R. To evaluate the sensitivity of T2-DAG with respect to the mis-specification of R, we conducted additional simulations with two continuous confounders. Specifically, we generated two groups of gene expression data accord-
Figure 4: Empirical type I error rates ($\delta = 0$) and powers ($|\delta| > 0$) of the various test assuming two continuous confounders, with the number of children nodes in the DAG set to $p_0 = 0.8p$, and the number of non-zero signals set to $q = 0.5p$. Please note that the Hotelling’s $T^2$ test (T2) is not presented in columns 2 and 3 because it is not applicable when $n_1 + n_2 \leq p$.

Figure 4: Empirical type I error rates ($\delta = 0$) and powers ($|\delta| > 0$) of the various test assuming two continuous confounders, with the number of children nodes in the DAG set to $p_0 = 0.8p$, and the number of non-zero signals set to $q = 0.5p$. Please note that the Hotelling’s $T^2$ test (T2) is not presented in columns 2 and 3 because it is not applicable when $n_1 + n_2 \leq p$.

Figure 4: Empirical type I error rates ($\delta = 0$) and powers ($|\delta| > 0$) of the various test assuming two continuous confounders, with the number of children nodes in the DAG set to $p_0 = 0.8p$, and the number of non-zero signals set to $q = 0.5p$. Please note that the Hotelling’s $T^2$ test (T2) is not presented in columns 2 and 3 because it is not applicable when $n_1 + n_2 \leq p$.

ing to $x^{(l)}_i = (I - Q^\top)^{-1}(C^\top w_i + \epsilon_i) + \mu_l$, $l = 1, 2$, $i = 1, \ldots, n_l$. Here, $\mu_1$, $\mu_2$, $Q$ and $\epsilon_i$ were generated following the same procedure described in Section 3.1. The confounding variables $w_i = (w_{i,1}, w_{i,2})^\top$, $i = 1, \ldots, n_1 + n_2$, were generated independently from $\mathcal{N}(0, \Sigma_w)$ with $\Sigma_w = \text{diag}((0.2)^2/\|Q\|_{\text{max}})$. Twenty percent of the entries in each row of $C$ were generated independently from $\mathcal{N}(0, \sqrt{0.2})$, and the remaining entries were set to 0. As such, a random set of $0.2p$ genes were conditionally associated with each confounder. For T2-DAG test, we assumed that the confounders $w_i$ were latent, and T2-DAG was still implemented based on model (2).

We observe from Fig. 4 that all tests have well-controlled type I error rates except for GCT, which shows high inflation in type I error rates. When the signals are not too weak (i.e., $|\delta| > 0.05$), T2-DAG is more powerful than all other tests implemented. Additional simulation results based on $n_1 = n_2 = 50$, $p = 50$ and $n_1 = n_2 = 100$, $p = 100$ have similar patterns and are summarized in Fig. S7. These results show that T2-DAG can still outperform the existing tests even with unadjusted confounders.
Figure 5: Empirical type I error rates ($\delta = 0$) and powers ($|\delta| > 0$) of the various tests under $n_1 = n_2 = 50$ and $p = 100$ with error terms generated from log-normal distributions based on the raw data (row 1) or log-transformed data (row 2). The number of children nodes in the DAG is set to $p_0 = 0.6p$ (columns 1 and 2) or $p_0 = 0.8p$ (columns 3 and 4), and the number of non-zero signals is set to $q = 0.1p$ (columns 1 and 3) or $0.5p$ (columns 2 and 4). Additional results under different settings of $n_1$, $n_2$, $p$ and results assuming error terms of gamma or uniform distributions are summarized in Figs. S8 - S18.

3.5 Results given non-Gaussian error terms

The theory of T2-DAG requires the assumption of Gaussian error terms, which can be violated in reality. We therefore conducted an additional sensitivity analysis on T2-DAG where we considered three alternative distributions: log-normal distribution, gamma distribution, and uniform distribution. Detailed simulation settings are summarized in Section 2.4 of the supplementary material. We can observe from the first row of Fig. 5 and Figs. S8 - S18 that T2-DAG tests perform similarly as in Section 3.2 with well-controlled type I error rates and higher powers than the existing tests, showing the robustness of T2-DAG to these distributions of the error terms.

We further conducted a log transformation on the data under the scenarios of log-normal and gamma error terms where the simulated gene expression data were right skewed. Interestingly, this additional data transformation step lead to a substantial gain in power for all tests including
T2-DAG in all scenarios, except for GCT in a few scenarios where it had powers lower than 0.15 under all signal strengths (Figs. 5 and S8 - S14, bottom row). Furthermore, while there is no obvious change in the type I error rates of the existing tests, T2-DAG consistently has a similar or more strict type I error control after the log transformation. These findings suggest that for right skewed gene expression data which is commonly seen in genetics studies we can conduct a log transformation to improve the performance of T2-DAG.

4 A real data example

In lung cancer diagnosis and prognosis, extensive research has been conducted utilizing gene expression profiling to identify lung cancer-related gene pathways for therapeutic research (Weiss and Kingsley, 2008; Shao et al., 2010; Cai and Jiang, 2014; Dancik and Theodorescu, 2014; Chang et al., 2015; Long et al., 2019; Venugopal et al., 2019; Kong et al., 2020). In this application, we aim to identify differentially expressed gene pathways between different stages of lung cancer, the findings of which could provide insights into the underlying genetic mechanisms of lung cancer progression.

We applied T2-DAG $Z$ and T2-DAG $\chi^2$, as well as the existing tests considered in the simulation studies, to a gene expression dataset for lung cancer patients in The Cancer Genome Atlas (TCGA) Program (Network et al., 2014). The dataset was obtained from the Lung Cancer Explorer (LCE) (Quantitative Biomeical Research Center, 2019) with standard quality control, where the expression values of zero were set to the overall minimum value and all data were then $\log_2$-transformed so that the gene expression data are approximately normal for all lung cancer stages (Cai et al., 2014, 2019). Expression levels of 20429 genes were measured for lung cancer tissue samples collected from 513 patients, among whom $N_1 = 278$, $N_2 = 124$, $N_3 = 84$ and $N_4 = 27$ were diagnosed with stage I, II, III and IV lung cancer, respectively. Additionally, $N_0 = 59$ normal tissue samples were collected from 59 of the 513 patients. When comparing
cancer tissues with normal tissues, we excluded the tumor samples of these 59 patients to avoid dependent samples.

We considered $H = 206$ common human pathways related to signaling, metabolic and human diseases from the KEGG database (Kanehisa and Goto, 2000; Kanehisa et al., 2019). Pathway information on previously identified gene interactions is categorized into activation/expression or inhibition/repression of one gene by another (Kanehisa and Goto, 2000; Kanehisa et al., 2019); see Fig. 1 for an illustrative example. Detailed characteristics of the pathways are summarized in Table S1. Among the 206 pathways, 34 have circles (including self loops) in the corresponding graphs, among which 18 has 1 circle only, 8 have 2 circles, 4 have 3 circles, and 4 have more than 3 circles each. We removed these circles when implementing the proposed T2-DAG test but used the complete graph information when implementing the Graph T2 test. Other than this, all the tests were implemented in the same way as in Section 3. We further applied Bonferroni correction (Bonferroni, 1936) with a significance level $\alpha = \alpha_0 / H$ for each test to control the family-wise error rate (FWER) at $\alpha_0 = 0.05$.

All tests except aSPU conclude that all 206 gene pathways are differentially expressed between normal tissues and tissues of each cancer stage (Tables S7 - S10). Table 1 summarizes the number of differentially expressed gene pathways detected by each test between stage I and stage III lung cancer and between stage II and stage III lung cancer. For the comparison between stage I and stage III lung cancer, the sample sizes ($N_1 = 278$, $N_2 = 124$) are large relative to the corresponding $p$, $p_0$ and $d$ for all pathways. The Hotelling’s $T^2$ test and Graph T2 test identify 8 and 10 significant pathways, respectively, while CH-Q, aSPU, GCT, SK and ARHT identify much more pathways (i.e., 96, 95, 82, 59 and 37, respectively). On the other hand, the CLX test, which is powerful for detecting sparse signals, only identifies 1 significant pathway. This may indicate that the signals of differential gene expression in real pathways tend to be dense, i.e., scatter among many correlated genes. The existing tests in total identify 129 unique pathways, while the T2-DAG $\chi^2$ test and $Z$ test identify 127 and 144 significant pathways, respectively, which cover
Table 1: Results of the hypothesis tests on 206 KEGG pathways for differential gene expression between stage I and stage III lung cancer (top panel, $N_1 = 278$, $N_3 = 84$) and between stages II and stage III lung cancer (bottom panel, $N_2 = 124$, $N_3 = 84$). In each panel, the numbers on the diagonal line are the numbers of significant pathways identified by the corresponding methods; the numbers in the $(i,j)$-th cell in the upper and lower triangle are the number of significant pathways detected by both method $i$ and method $j$ and the estimated correlation of $-\log_{10}(p)$ between the two methods, respectively. “T2” denotes the Hotelling’s $T^2$ test.

|                | Stage I versus Stage III | Stage II versus Stage III |
|----------------|--------------------------|---------------------------|
|                | T2-DAG $\chi^2$         | T2-DAG Z                  | T2-DAG $\chi^2$         | T2-DAG Z                  |
| T2-DAG $\chi^2$ | 127                      | 127                       | 0                        | 0                        |
| T2-DAG Z        | 0.98                     | 1.44                      | 0.21                     | 0.21                     |
| Graph T2        | 0.43                     | 0.45                      | 0.42                     | 0.42                     |
| T2             | 0.23                     | 0.23                      | 0.34                     | 0.34                     |
| CH-Q            | 0.57                     | 0.59                      | 0.48                     | 0.48                     |
| SK              | 0.75                     | 0.77                      | 0.48                     | 0.48                     |
| CLX             | 0.54                     | 0.56                      | 0.34                     | 0.34                     |
| GCT             | 0.81                     | 0.76                      | 0.36                     | 0.36                     |
| aSPU            | 0.51                     | 0.55                      | 0.27                     | 0.27                     |
| ARHT            | 0.56                     | 0.56                      | 0.26                     | 0.26                     |

115 and 122, respectively, of the 129 pathways identified by the existing tests. Furthermore, the $-\log_{10}$-transformed $p$-values ($-\log_{10}(p)$) of the T2-DAG tests are positively correlated with those of all the other tests; in particular, results of T2-DAG tests are in line with those of CH-Q, SK, CLX, GCT, aSPU and ARHT, with correlations higher than 0.5. The two T2-DAG tests yield highly correlated results (the correlation of $-\log_{10}(p)$ equals 0.98). T2-DAG Z detects a slightly larger number of significant pathways than T2-DAG $\chi^2$, and for the pathways detected by both tests, T2-DAG Z tends to give smaller $p$-values than T2-DAG $\chi^2$. This is different from what was observed from simulations where T2-DAG Z tends to have lower rejection rates than T2-DAG.
\(\chi^2\), which could be due to the difference in data structures between the simulated data and real data. This inconsistency in relative performance is observed not only for the two T2-DAG tests, but also for other tests such as CH-Q and ARHT, which perform similarly on the simulated data but yield quite different results on the lung cancer data where they detect 96 and 37 differentially expressed pathways, respectively, between stage I and stage III lung cancer.

For the comparison between stage II and stage III lung cancer, all tests except GCT and aSPU detect no differentially expressed gene pathway. This may be due to the limited sample sizes \((N_3 = 124, N_4 = 84)\) and/or relatively small differences between stage II and stage III lung cancer. On the other hand, aSPU identifies 1 significant pathway, while GCT identifies 21 significant pathways that do not include the one identified by aSPU. These 21 pathways are possibly false positives because of two reasons. First, none of the pathways is detected by any other test. Second, in our simulation studies, GCT could have substantially inflated type I error rates under small \(p\) settings (Fig. 2, Fig. 4, Fig. S1 and Fig. S7).

Figure 6: \(-\log_{10}(p)\) of the various tests for differential expression between stage I and stage III lung cancer on 20 randomly selected human pathways.

To take a closer look at these results, we randomly selected 20 of the 206 pathways and report the corresponding \(-\log_{10}(p)\)s (Fig. 6: stage I versus stage III; Fig. S18: stage II versus stage III). Overall, the \(p\)-values for comparison between stage I and stage III lung cancer show consistency
across different tests. Specifically, for pathways on which all tests yield insignificant results with large p-values, the T2-DAG tests give similar and sometimes even larger p-values compared to the other tests; see the Sulfur relay system pathway and the Circadian rhythm pathway in Fig. 6. For pathways on which T2-DAG yields highly significant results with the smallest p-values, many of the other tests also yield significant results; see the Calcium signaling pathway and the Type I diabetes mellitus in Fig. 6. On the contrary, for the pathways identified by GCT to be significantly differentially expressed between stage II and stage III cancer, all the other tests yield insignificant results with large p-values (Fig. 7).

We also compared other pairs of cancer stages and summarized the results in Section 3 of the supplementary material. Results for the comparison of stage I versus II, stage I versus IV, and stage II versus IV cancer have similar patterns as the comparison between stage I and III cancer. Similar to the comparison between stage II and III cancer, when comparing stage III to stage IV cancer, none of the tests identifies any significant pathway except for GCT which identifies 5 significant pathways. Detailed results for all the tests are summarized in Tables S1 - S10.

Figure 7: The $-\log_{10}(p)$ of the comparison between stage II and stage III lung cancer ($N_2 = 124$, $N_3 = 84$) for the pathways identified by any test to be significantly differentially expressed between the two cancer stages.
5 Discussion

The proposed T2-DAG $Z$ test and the alternative T2-DAG $\chi^2$ test have illustrated their advantages over the existing tests through our simulation studies and application to a lung cancer dataset. When establishing asymptotic properties, the T2-DAG $Z$ test is favored over the alternative T2-DAG $\chi^2$ test since it requires less stringent conditions and thus is presented as our main result. Although facing theoretical limitations, the alternative T2-DAG $\chi^2$ test has excellent finite-sample performance where it has slightly higher but still well-controlled type I error rates and higher powers than T2-DAG $Z$, as shown by simulations. In the lung cancer application, however, T2-DAG $\chi^2$ appears to be slightly more conservative than T2-DAG $Z$ but still identifies a much larger number of differentially expressed pathways than the existing tests for stage I versus stage II, III, or IV cancer. In applications, the choice between T2-DAG $Z$ and T2-DAG $\chi^2$ can be made based on the theoretical properties and simulated finite-sample results under the specific data scenarios of sample size ($n_1$ and $n_2$), dimension ($p$), and sparsity of gene interactions ($p_0$ and $d$) inferred from the auxiliary pathway information.

The proposed T2-DAG test was motivated by gene pathway analyses, but it is generally applicable to problems in many other areas where similar auxiliary graph information is available. For example, in human microbiome studies for identifying host-microbe associations, the auxiliary graph information is provided by the phylogenetic tree that characterizes the evolutionary relationships among the microbes.

We have illustrated the robustness of T2-DAG with respective to model mis-specifications due to missing/inaccurate edge information or unadjusted confounding effects. While our simulation results are encouraging in these scenarios, the current estimation procedure may no longer be optimal. Thus, we consider the following two extensions of T2-DAG as our future work. First, while T2-DAG requires that the auxiliary pathway information can be represented by a DAG, it is feasible to extend it to non-DAG pathways with more complex gene interactions.
For example, the recently proposed linear SEM for cyclic mixed graphs (Améndola et al., 2020) shows a possible way to account for feedback loops and undirected edges in the gene pathway. Second, the proposed test assumes a diagonal covariance matrix $R$ for the error term $\epsilon$ in (2). As discussed in Section 3.4, this assumption may fail due to latent confounders. In this case, we may consider a non-diagonal but sparse $R$, i.e., only a small number of off-diagonal entries of $R$ are non-zero, and estimate $R$ via graphical lasso (Friedman et al., 2008) or other penalized estimation approaches for high-dimensional precision matrix (Zhang and Zou, 2014; Kuismin et al., 2017; Fan et al., 2019). Furthermore, when measurements are available for some potential confounders, we can directly incorporate them into the SEM as covariates. We have in fact observed promising performance of this strategy through additional simulation studies, but the theoretical properties of the strategy require further investigation.

While the current T2-DAG test only incorporates information on the presence/absence of gene interactions, it is possible to further incorporate the signs of the interactions, such as activation (+) or inhibition (−) of one gene by another. One possible strategy is to conduct modeling in a Bayesian framework, where the signs of gene interactions can be accounted for by inducing truncated normal priors on the corresponding parameters in the coefficient matrix $Q$. The asymptotic properties of the test, however, need to be re-investigated.

Finally, we note that the key component of the proposed T2-DAG test is the novel DAG-informed estimator for the precision matrix (see Equations (3) and (4)), which could provide substantially improved power through incorporating external pathway information that describes gene interactions. Besides being incorporated into our T2-DAG test, this novel estimator has plenty of other potential usage. For example, as illustrated in Section 4, T2-DAG, with a Hotelling’s $T^2$-type test statistic, is powerful against “dense” alternatives which are usually true for differentially expressed gene pathways. In applications where the signals are sparse, one may instead consider combining the proposed DAG-informed estimator with tests that are powerful against “sparse” alternatives, such as the CLX test (Cai et al., 2014), to further boost power. An-
other example is that the proposed DAG-informed estimator can be used for high-dimensional linear discriminant analysis (LDA) (Bouveyron et al., 2007) to improve the classification accuracy. Specifically, consider the Gaussian case where one wants to classify a random vector $x$ drawn with equal probability from one of the two Gaussian distributions, $\mathcal{N}(\mu_1, \Sigma)$ (class 1) and $\mathcal{N}(\mu_2, \Sigma)$ (class 2). When $\mu_1$, $\mu_2$ and $\Sigma$ are known, Fisher’s LDA rule, given by

$$C(x) = \begin{cases} 
1, & \delta^\top \Omega (x - \frac{\mu_1 + \mu_2}{2}) < 0 \\
2, & \delta^\top \Omega (x - \frac{\mu_1 + \mu_2}{2}) \geq 0,
\end{cases}$$

where $\delta = \mu_2 - \mu_1$, and $\Omega = \Sigma^{-1}$ denotes the precision matrix, is known to be optimal. In practice, one needs to estimate $\mu_1$, $\mu_2$, and $\Omega$. The proposed DAG-informed estimator of $\Omega$ can thus be applied here to facilitate the classification.

Conflcts of Interest

The authors declare no conflict of interest.

Data Availability

The data underlying this article are available at https://github.com/Jin93/T2DAG.

Supplementary Material

Supplementary material available at Biointformatics online includes technical results and proofs, detailed information on the pathways and gene expression data used in the data analysis, detailed simulation settings in the scenario of non-Gaussian error terms, as well as additional results on the simulated and real datasets.

29
Software

The R (R Development Core Team, 2021) package T2DAG which implements the proposed T2-DAG test is available on Github at https://github.com/Jin93/T2DAG.

References

C. Améndola, P. Dettling, M. Drton, F. Onori, and J. Wu. Structure learning for cyclic linear causal models. In Conference on Uncertainty in Artificial Intelligence, pages 999–1008. PMLR, 2020.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al. Gene ontology: tool for the unification of biology. Nature genetics, 25(1):25–29, 2000.

Z. Bai and H. Saranadasa. Effect of high dimension: by an example of a two sample problem. Statistica Sinica, pages 311–329, 1996.

C. Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerziali di Firenze, 8:3–62, 1936.

C. Bouveyron, S. Girard, and C. Schmid. High-dimensional discriminant analysis. Communications in Statistics—Theory and Methods, 36(14):2607–2623, 2007.

B. Cai and X. Jiang. Revealing biological pathways implicated in lung cancer from tcga gene expression data using gene set enrichment analysis. Cancer Informatics, 13:CIN–S13882, 2014.

L. Cai, S. Lin, L. Girard, Y. Zhou, L. Yang, B. Ci, Q. Zhou, D. Luo, B. Yao, H. Tang, et al. Lce: an open web portal to explore gene expression and clinical associations in lung cancer. Oncogene, 38(14):2551–2564, 2019.
T. T. Cai, W. Liu, and Y. Xia. Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):349–372, 2014.

J. T.-H. Chang, Y. M. Lee, and R. S. Huang. The impact of the cancer genome atlas on lung cancer. *Translational Research*, 166(6):568–585, 2015.

L. S. Chen, D. Paul, R. L. Prentice, and P. Wang. A regularized hotelling’s t 2 test for pathway analysis in proteomic studies. *Journal of the American Statistical Association*, 106(496):1345–1360, 2011.

S. X. Chen and Y.-L. Qin. A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*, 38(2):808–835, 2010.

H. Chun, X. Zhang, and H. Zhao. Gene regulation network inference with joint sparse gaussian graphical models. *Journal of Computational and Graphical Statistics*, 24(4):954–974, 2015.

G. O. Consortium. The gene ontology resource: 20 years and still going strong. *Nucleic acids research*, 47(D1):D330–D338, 2019.

G. M. Dancik and D. Theodorescu. Robust prognostic gene expression signatures in bladder cancer and lung adenocarcinoma depend on cell cycle related genes. *PloS one*, 9(1):e85249, 2014.

R. Fan, B. Jang, Y. Sun, and S. Zhou. Precision matrix estimation with noisy and missing data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2810–2819. PMLR, 2019.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.
K. B. Gregory, R. J. Carroll, V. Baladandayuthapani, and S. N. Lahiri. A two-sample test for equality of means in high dimension. *Journal of the American Statistical Association*, 110(510):837–849, 2015.

H. Hotelling. The generalization of student’s ratio. *The Annals of Mathematical Statistics*, 2(3):360–378, 1931.

J. Z. Huang, N. Liu, M. Pourahmadi, and L. Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98, 2006.

L. Jacob, P. Neuvial, and S. Dudoit. More power via graph-structured tests for differential expression of gene networks. *The Annals of Applied Statistics*, 6(2):561–600, 2012.

M. Kanehisa and S. Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

M. Kanehisa, Y. Sato, M. Furumichi, K. Morishima, and M. Tanabe. New approach for understanding genome variations in kegg. *Nucleic acids research*, 47(D1):D590–D595, 2019.

C. Kong, Y.-X. Yao, Z.-T. Bing, B.-H. Guo, L. Huang, Z.-G. Huang, and Y.-C. Lai. Dynamical network analysis reveals key micrornas in progressive stages of lung cancer. *PLOS Computational Biology*, 16(5):e1007793, 2020.

A. Krishnamoorthy and D. Menon. Matrix inversion using cholesky decomposition. In *2013 signal processing: Algorithms, architectures, arrangements, and applications (SPA)*, pages 70–72. IEEE, 2013.

M. Kuismin, J. Kemppainen, and M. Sillanpää. Precision matrix estimation with rope. *Journal of Computational and Graphical Statistics*, 26(3):682–694, 2017.

B. Li, H. Chun, and H. Zhao. Sparse estimation of conditional graphical models with application to gene networks. *Journal of the American Statistical Association*, 107(497):152–167, 2012.
H. Li, A. Aue, D. Paul, J. Peng, and P. Wang. An adaptable generalization of hotelling’s $t^2$ test in high dimension. *The Annals of Statistics*, 48(3):1815–1847, 2020.

P.-L. Loh and P. Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *The Journal of Machine Learning Research*, 15(1):3065–3105, 2014.

T. Long, Z. Liu, X. Zhou, S. Yu, H. Tian, and Y. Bao. Identification of differentially expressed genes and enriched pathways in lung cancer using bioinformatics analysis. *Molecular medicine reports*, 19(3):2029–2040, 2019.

M. Lopes, L. Jacob, and M. J. Wainwright. A more powerful two-sample test in high dimensions using random projection. In *Advances in Neural Information Processing Systems*, pages 1206–1214, 2011.

M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014.

C. G. A. R. Network et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543–550, 2014.

D. Nishimura. Biocarta. *Biotech Software & Internet Report*, 2(3):117–120, 2001. doi: 10.1089/152791601750294344. URL https://doi.org/10.1089/152791601750294344.

J. Peters and P. Bühlmann. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

Quantitative Biomeical Research Center. Lung Cancer Explorer. http://lce.biohpc.swmed.edu/, 2019. Accessed: 2021-06-15.

M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.
W. Shao, D. Wang, and J. He. The role of gene expression profiling in early-stage non-small cell lung cancer. *Journal of thoracic disease*, 2(2):89, 2010.

A. Shojaie and G. Michailidis. Analysis of gene sets based on the underlying regulatory network. *Journal of Computational Biology*, 16(3):407–426, 2009.

A. Shojaie and G. Michailidis. Network enrichment analysis in complex experiments. *Statistical applications in genetics and molecular biology*, 9(1), 2010.

M. S. Srivastava and M. Du. A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis*, 99(3):386–402, 2008.

M. S. Srivastava and T. Kubokawa. Tests for multivariate analysis of variance in high dimension under non-normality. *Journal of Multivariate Analysis*, 115:204–216, 2013.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

N. Venugopal, J. Yeh, S. K. Kodeboyina, T. J. Lee, S. Sharma, N. Patel, and A. Sharma. Differences in the early stage gene expression profiles of lung adenocarcinoma and lung squamous cell carcinoma. *Oncology Letters*, 18(6):6572–6582, 2019. doi: 10.3892/ol.2019.11013.

G. J. Weiss and C. Kingsley. Pathway targets to explore in the treatment of non-small cell lung cancer. *Journal of Thoracic Oncology*, 3(11):1342–1352, 2008.

A. Wille, P. Zimmermann, E. Vranová, A. Fürholz, O. Laule, S. Bleuler, L. Hennig, A. Prelić, P. von Rohr, L. Thiele, et al. Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. *Genome biology*, 5(11):R92, 2004.

W. B. Wu and M. Pourahmadi. Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90(4):831–844, 2003.
G. Xu, L. Lin, P. Wei, and W. Pan. An adaptive two-sample test for high-dimensional means.

*Biometrika*, 103(3):609–624, 2016.

T. Zhang and H. Zou. Sparse precision matrix estimation via lasso penalized d-trace loss.

*Biometrika*, 101(1):103–120, 2014.
Supplementary Materials for “T2-DAG: a powerful test for differentially expressed gene pathways via graph-informed structural equation modeling”

Jin Jin, Yue Wang

1 Proof of Theorems

In this proof, we consider the high-dimensional regime where \((p, d, p_0) \to \infty\) as \((n_1, n_2) \to \infty\).

We begin with some additional notations. For a \(p\)-dimensional random vector \(\mathbf{v} = (v_1, \ldots, v_p)^T\), we define the following vector norms,

\[
\|\mathbf{v}\|_q := \left( \sum_{j=1}^{p} |v_j|^q \right)^{1/q} \quad \text{for } 0 < q < \infty, \quad \|\mathbf{v}\|_\infty := \max_{1 \leq j \leq p} |v_j|.
\]

For a \(p \times p\) matrix \(\mathbf{M} = (m_{jk})_{j,k=1,\ldots,p}\), we define

\[
\|\mathbf{M}\|_q := \max_{\|\mathbf{v}\|_q = 1} \|\mathbf{Mv}\|_q, \quad \|\mathbf{M}\|_{\max} := \max_{(j,k)} |m_{jk}|, \quad \|\mathbf{M}\|_F := \left( \sum_{j,k} |m_{jk}|^2 \right)^{1/2},
\]

for \(0 < q < \infty\). We use \(I_p\) to denote a \(p \times p\) identity matrix. For any index set \(S \subseteq \{1, \ldots, p\}\), let \(\mathbf{v}_S\) denote the subvector of \(\mathbf{v}\) with entries indexed by \(S\). Similarly, \(\mathbf{M}_{S,S}\) denotes the submatrix of \(\mathbf{M}\) with rows and columns indexed by \(S\).

We first cite the following result (Theorem 6.1, Wainwright, 2019).

**Lemma 1.1.** Let \(\mathbf{x}_1, \ldots, \mathbf{x}_n\) be \(p \times 1\) random vectors drawn i.i.d from a multivariate normal distribution with mean \(0\) and covariance \(\Sigma\), and \(\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_n]^T\). Let \(\sigma_{\max}(\mathbf{X}), \sigma_{\min}(\mathbf{X})\),
\( \gamma_{\max}(\Sigma) \) and \( \gamma_{\min}(\Sigma) \), respectively, denote the largest and smallest singular value of \( X \) and the largest and smallest eigenvalue of \( \Sigma \). For all \( \delta > 0 \), we then have

\[
P \left[ \frac{\sigma_{\max}(X)}{\sqrt{n}} \geq \sqrt{\gamma_{\max}(\Sigma)}(1 + \delta) + \sqrt{\frac{\text{tr}(\Sigma)}{n}} \right] \leq e^{-n\delta^2/2},
\]

and

\[
P \left[ \frac{\sigma_{\min}(X)}{\sqrt{n}} \leq \sqrt{\gamma_{\min}(\Sigma)}(1 - \delta) - \sqrt{\frac{\text{tr}(\Sigma)}{n}} \right] \leq e^{-n\delta^2/2}.
\]

Recall that \( S_j = \{ i : a_{ij} = 1 \} \), where \( a_{ij} \) is the \((i, j)\)-th entry of the adjacency matrix \( A \), and \( \hat{\Sigma} = (n_1 + n_2)^{-1} \left( \sum_{k=1}^{n_1} (x_k^{(1)} - \bar{x}^{(1)}) \otimes^2 + \sum_{k=1}^{n_2} (x_k^{(2)} - \bar{x}^{(2)}) \otimes^2 \right) \), where for any vector \( a \), \( a \otimes^2 = aa^\top \).

The next result bounds the spectral norm distance between \( \hat{\Sigma}_{S_j, S_j} \) and \( \Sigma_{S_j, S_j} \) for all \( j \in \mathcal{J} \), where \( \mathcal{J} = \{ j : |S_j| > 0 \} \).

**Lemma 1.2.** As \( n_1, n_2 \to \infty \), we have

\[
\max_{j \in \mathcal{J}} \frac{\| \Sigma_{S_j, S_j}^{1/2} \hat{\Sigma}_{S_j, S_j} \Sigma_{S_j, S_j}^{1/2} - I_{|S_j|} \|_2}{|S_j|} = O_p \left( \frac{\log p_0}{n_1 + n_2} \right).
\]

Furthermore, if \( d \log p_0 = o(n_1 + n_2) \), then

\[
\max_{j \in \mathcal{J}} \frac{\| \Sigma_{S_j, S_j}^{1/2} \hat{\Sigma}_{S_j, S_j}^{-1} \Sigma_{S_j, S_j}^{1/2} - I_{|S_j|} \|_2}{|S_j|} = O_p \left( \frac{\log p_0}{n_1 + n_2} \right).
\]

**Proof of Lemma 1.2.** Recalling that

\[
z_i = \begin{cases} 
    x_k^{(1)} - \mu_1 & k = 1, 2, \ldots, n_1 \\
    x_k^{(2)} - \mu_2 & k = n_1 + 1, \ldots, n_1 + n_2
\end{cases},
\]

\[
|S_j| = \{ i : x_i = 1 \}
\]

and

\[
\Sigma_{S_j, S_j} = \sqrt{\Sigma} S_j \Sigma S_j \sqrt{\Sigma}
\]

for any symmetric \( n \times n \) matrix \( \Sigma \).
we write

\[ \hat{\Sigma} = \frac{1}{n_1 + n_2} \sum_{k=1}^{n_1+n_2} z_k^{\otimes 2} + \frac{n_1}{n_1 + n_2} (\overline{x}^{(1)} - \mu_1)^{\otimes 2} + \frac{n_2}{n_1 + n_2} (\overline{x}^{(2)} - \mu_2)^{\otimes 2}. \]

Letting \( w_k = \Sigma^{-1/2} z_k \), it is easy to check that \( w_k \sim i.i.d. N(0, I_p) \). Thus, we have

\[
\left\| \Sigma_{S_j}^{-1/2} \hat{\Sigma}_{S_j} \Sigma_{S_j}^{-1/2} - I_{|S_j|} \right\|_2 \leq \left\| \frac{1}{n_1 + n_2} \sum_{k=1}^{n_1+n_2} (w_k^{\otimes 2})_{S_j} - I_{|S_j|} \right\|_2 + \frac{n_1}{n_1 + n_2} \left\| \sum_{S_{j'}}^1 (\overline{x}^{(1)} - \mu_1)_{S_{j'}} \right\|_2^2 + \frac{n_2}{n_1 + n_2} \left\| \sum_{S_{j'}}^1 (\overline{x}^{(2)} - \mu_2)_{S_{j'}} \right\|_2^2.
\]

Using Lemma 1.1, we know for all \( \delta > 0 \),

\[
P \left\{ \gamma_{\max} \left( \frac{1}{n_1 + n_2} \sum_{k=1}^{n_1+n_2} (w_k^{\otimes 2})_{S_j} \right) \geq 1 + \delta + \sqrt{\frac{|S_j|}{n_1 + n_2}} \right\} \leq e^{-\frac{(n_1+n_2)\delta^2}{2}}
\]

and

\[
P \left\{ \gamma_{\min} \left( \frac{1}{n_1 + n_2} \sum_{k=1}^{n_1+n_2} (w_k^{\otimes 2})_{S_j} \right) \leq 1 - \delta - \sqrt{\frac{|S_j|}{n_1 + n_2}} \right\} \leq e^{-\frac{(n_1+n_2)\delta^2}{2}}.
\]

These together yield

\[
P \left\{ \frac{1}{n_1 + n_2} \sum_{k=1}^{n_1+n_2} (w_k^{\otimes 2})_{S_j} - I_{|S_j|} \geq \left( \delta + \sqrt{\frac{|S_j|}{n_1 + n_2}} \right)^2 \right\} \leq e^{-\frac{(n_1+n_2)\delta^2}{2}}.
\]

By taking \( \delta = \sqrt{4|S_j| \log p_0/(n_1 + n_2)} \), we have

\[
P \left\{ |S_j|^{-1} \left\| \frac{1}{n_1 + n_2} \sum_{k=1}^{n_1+n_2} (w_k^{\otimes 2})_{S_j} - I_{|S_j|} \right\|_2 \geq \frac{1}{n_1 + n_2} \left( \sqrt{2 \log p_0} + 1 \right)^2 \right\} \leq p_0^{-2|S_j|}.
\]
Taking the union over \( j \in \mathcal{J} \), we have

\[
P \left\{ \max_{j \in \mathcal{J}} \left\{ |S_j|^{-1} \left\| \frac{1}{n_1 + n_2} \sum_{k=1}^{n_1+n_2} (w_{k_j^2}) S_{j|S_j} - I_{|S_j|} \right\|_2 \right\} \geq \frac{1}{n_1 + n_2} \left( \sqrt{2 \log p_0 + 1} \right)^2 \right\}
\]

\[
\leq \sum_{j \in \mathcal{J}} p_0^{-2|S_j|} \leq p_0^{-1},
\]

which goes to 0 as \( n_1, n_2 \to \infty \). This indicates that as \( n_1, n_2 \to \infty \), we have

\[
\max_{j \in \mathcal{J}} |S_j|^{-1} \left\| \frac{1}{n_1 + n_2} \sum_{k=1}^{n_1+n_2} w_{k_j} S_{j|S_j} - I_{|S_j|} \right\|_2 = O_p \left( \frac{\log p_0}{n_1 + n_2} \right).
\]

Also,

\[
\frac{n_1}{n_1 + n_2} \left\| \sum_{S_j,S_j} (\bar{v}(1) - \mu_1) S_j \right\|_2^2 = \frac{n_1}{n_1 + n_2} \left\| n_1^{-1} \sum_{k=1}^{n_1} (w_k S_j) \right\|_2^2
\]

Note that \( \left\| n_1^{-1/2} \sum_{k=1}^{n_1} (w_k S_j) \right\|_2^2 \sim \chi^2_{|S_j|} \), therefore by using the standard tail bound for \( \chi^2 \) variables, we can get

\[
P \left\{ \left\| n_1^{-1/2} \sum_{k=1}^{n_1} (w_k S_j) \right\|_2^2 > |S_j| n_1 t \right\} \leq 2e^{-|S_j| n_1 t/8} \leq 2e^{-n_1 t/8}, \text{ for all } t > \frac{1}{n_1}
\]

Hence, taking the union over \( j \in \mathcal{J} \) gives

\[
P \left\{ \max_{j \in \mathcal{J}} \left\| n_1^{-1/2} |S_j|^{-1} \sum_{k=1}^{n_1} (w_k S_j) \right\|_2^2 - 1 > n_1 t \right\} \leq e^{\log(2p_0) - n_1 t/8}, \text{ for all } t > \frac{1}{n_1}.
\]

\[\square\]
By taking $t = 16 \log(2p_0)/n_1 > n_1^{-1}$, we get
\[
P \left\{ \max_{j \in J} \left| |S_j|^{-1} \left| n_1^{-1/2} \sum_{k=1}^{n_1} (w_k)_j \right|^2 - 1 \right| > 16 \log(2p_0) \right\} \leq (2p_0)^{-1}.
\]
This indicates that
\[
\frac{n_1}{n_1 + n_2} \max_{j \in J} |S_j|^{-1} \left\| n_1^{-1/2} \sum_{k=1}^{n_1} (w_k)_j \right\|_2^2 = O_p \left( \frac{\log p_0}{n_1 + n_2} \right).
\]
Similarly, we can show that
\[
\frac{n_2}{n_1 + n_2} \left\| \Sigma_{S_j}^{-1/2} (\bar{x}(2) - \mu_j)_{S_j} \right\|_2^2 = O_p \left( \frac{\log p_0}{n_1 + n_2} \right). \tag{5}
\]
This completes the proof of (3) in Lemma 1.2. To show (4) in Lemma 1.2, we only need to note that
\[
\left\| \Sigma_{S_j}^{-1/2} \{ \hat{\Sigma}_{S_j,S_j} \}^{-1} \Sigma_{S_j,S_j} - I_{|S_j|} \right\|_2 \leq \frac{\left\| \Sigma_{S_j,S_j}^{-1/2} \hat{\Sigma}_{S_j,S_j} \Sigma_{S_j,S_j}^{-1/2} - I_{|S_j|} \right\|_2}{1 - \left\| \Sigma_{S_j,S_j}^{-1/2} \hat{\Sigma}_{S_j,S_j} \Sigma_{S_j,S_j}^{-1/2} - I_{|S_j|} \right\|_2};
\]
here, we use the matrix variant of Taylor expansion. It follows from (3) that if $d \log p_0 = o(n_1 + n_2)$, we then have $\left\| \Sigma_{S_j,S_j}^{-1/2} \{ \hat{\Sigma}_{S_j,S_j} \}^{-1} \Sigma_{S_j,S_j} - I_{|S_j|} \right\|_2 = O(\left\| \Sigma_{S_j,S_j}^{-1/2} \hat{\Sigma}_{S_j,S_j} \Sigma_{S_j,S_j}^{-1/2} - I_{|S_j|} \right\|_2)$. This yields (4).

For any set $E$, let $A_E = \{ x_{i,l} : l = 1, 2; i = 1, \ldots, n_1 \}$. The following lemma characterizes the properties of $\hat{Q}$ and $\hat{R}$.

**Lemma 1.3.** Under Assumptions 1-3, for $j \in J$ and any arbitrary set $E$ that satisfies $S_j \subseteq E \subseteq \{1, \ldots, j-1\}$, we have

1. Conditioned on $A_E$, $\hat{q}_{j,S_j} \perp \hat{r}_j$ and $(\bar{x}^{(1)} - \bar{x}^{(2)})^\top e_j \perp \hat{r}_j$, where $e_j \in \mathbb{R}^p$ denotes the vector where the $j$-th entry equals 1 and 0 elsewhere.

2a. Conditioned on $A_E$, $\hat{q}_{j,S_j}$ follows a multivariate normal distribution with mean $q_{j,S_j}$ and
covariance $r_j(n_1 + n_2)^{-1} \{ \hat{\Sigma}_{S_j,S_j} \}^{-1}$.

(2b) Conditioned on $AE$, $\hat{q}_{j,S_j} \perp (\bar{x}^{(1)} - \bar{x}^{(2)})^\top e_j$.

(2c) For $k < j$, we have

$$\mathbb{E} \left[ (\hat{q}_{j,S_j} - q_{j,S_j})(\hat{q}_{k,S_k} - q_{k,S_k})^\top | A_{S_j \cup S_k \cup \{k\}} \right] = 0_{|S_j| \times |S_k|}.$$  

(3a) For $j = 1, \ldots, p$, we have $\mathbb{E} [\hat{r}_j^{-1} | A_E] = r_j^{-1}$ and $\text{Var}[\hat{r}_j^{-1} | A_E] = r_j^{-2}(n_1 + n_2 - |S_j| - 6)^{-1}$.

(3b) For $k < j$,

$$\mathbb{E} \left[ (\hat{r}_j^{-1} - r_j^{-1})(\hat{r}_k^{-1} - r_k^{-1}) | A_{S_j \cup S_k \cup \{k\}} \right] = 0.$$

Proof. We first recall some notations. We denote by $1_{n_1+n_2}$ the $(n_1 + n_2)$-dimensional vector of all ones and let $g = (g_1, \ldots, g_{n_1+n_2})^\top$ denote the group indicator; that is, $g_i$ equals 1 for $i = 1, \ldots, n_1$ and equals 0 for $i = n_1 + 1, \ldots, n_1 + n_2$. Then, letting $X = \left[ x_1^{(1)} \cdots x_{n_1}^{(1)} x_1^{(2)} \cdots x_{n_2}^{(2)} \right]^\top$, we consider

$$\|X_j - \theta_1 j 1_{n_1+n_2} - \theta_2 g - X_{S_j} q_{j,S_j}\|_2^2. \tag{6}$$

Here, $X_j$ denotes the $j$-th column of $X$; $X_{S_j}$ denotes the sub-matrix of $X$ with columns indexed by $S_j$; $q_{j,S_j}$ denotes the sub-vector of $q_j$ with entries indexed by $S_j$. Thus, letting $W = [1_{n_1+n_2} \ g]$, we obtain the OLS estimator of $q_{j,S_j}$ and $(\theta_1 j, \theta_2)^\top$:

$$\hat{q}_{j,S_j} = \left( X_{S_j}^\top (I_{n_1+n_2} - P_W) X_{S_j} \right)^{-1} X_{S_j}^\top (I_{n_1+n_2} - P_W) X_j,$$

$$\hat{\theta}_{1j, \hat{\theta}_{2j}}^\top = (W^\top W)^{-1} W^\top (X_j - X_{S_j} \hat{q}_{j,S_j}). \tag{7}$$

where $P_W = W(W^\top W)^{-1} W^\top$ is the orthogonal projection matrix onto the column space of $W$.  

6
Then, the “plug-in” estimator of \( r_j \) is

\[
\hat{r}_j = \frac{1}{n_1 + n_2 - |S_j| - 2} \left\| X_j - \hat{\theta}_{1j} 1_{n_1+n_2} - \hat{\theta}_{2j} g - X_{S_j} \hat{q}_{j,s_j} \right\|^2;
\]

for shorthand notation, let \( \hat{e}_j = X_j - \hat{\theta}_{1j} 1_{n_1+n_2} - \hat{\theta}_{2j} g - X_{S_j} \hat{q}_{j,s_j} \).

(1) Note that under Gaussianity, conditioned on the predictors, the OLS coefficient estimator is independent of the OLS variance estimator (note that the group indicator \( g_i \) is a fixed quantity). Thus, \( \hat{q}_{j,s_j} \perp \hat{r}_j \) given \( A_{S_j} \). We next prove \( \text{Cov} (\hat{e}_j, \hat{x}_j^{(l)} | A_{S_j}) = 0, \ l = 1, 2. \)

We first consider \( l = 1 \). By (6), we get

\[
\hat{e}_j = \{ I_{n_1+n_2} - P_W (I_{n_1+n_2} - X_{S_j} F) - X_{S_j} F \} X_j,
\]

where \( F = \left( X_{S_j}^\top (I_{n_1+n_2} - P_W) X_{S_j} \right)^{-1} X_{S_j}^\top (I_{n_1+n_2} - P_W) \). Then, since \( \hat{x}_j^{(l)} = n_1^{-1} g^\top X_j \), we have

\[
\text{Cov} (\hat{e}_j, \hat{x}_j^{(l)} | A_{S_j}) = n_1^{-1} \left\{ I_{n_1+n_2} - P_W (I_{n_1+n_2} - X_{S_j} F) - X_{S_j} F \right\} \text{Cov} (X_j | A_{S_j}) g.
\]

Note that \( X_j - \theta_{1j} 1_{n_1+n_2} - \theta_{2j} g - X_{S_j} q_{j,s_j} = \hat{e}_j \), where \( \hat{e}_j = (\epsilon_{1j}, \ldots, \epsilon_{n_1+n_2,j})^\top \). Thus,

\[
\text{Cov} (X_j | A_{S_j}) = \text{Cov} (\hat{e}_j) = r_j I_{n_1+n_2}.
\]
This leads to

\[
\text{Cov} \left( \tilde{\epsilon}_j, \tilde{x}_j^{(1)} \big| A_{S_j} \right) = n_1^{-1} r_j \left\{ I_{n_1+n_2} - \mathcal{P}_W(I_{n_1+n_2} - X_{S_j} F) - X_{S_j} F \right\} g = 0,
\]

because \( \mathcal{P}_W g = g \). For \( l = 2 \), we notice \( \tilde{x}_j^{(2)} = n_2^{-1}(1_{n_1+n_2} - g)^\top X_j \). Using similar tricks, we get

\[
\text{Cov} \left( \tilde{\epsilon}_j, \tilde{x}_j^{(2)} \big| A_{S_j} \right) = n_2^{-1} r_j \left\{ I_{n_1+n_2} - \mathcal{P}_W(I_{n_1+n_2} - X_{S_j} F) - X_{S_j} F \right\} (1_{n_1+n_2} - g) = 0,
\]

because \( \mathcal{P}_W(1_{n_1+n_2} - g) = 1_{n_1+n_2} - g \). Therefore, we have \( \tilde{\epsilon}_j \perp \tilde{x}_j^{(1)} - \tilde{x}_j^{(2)} \) given \( A_{S_j} \).

To generalize this result to any arbitrary set \( E \) that satisfies \( S_j \subseteq E \subseteq \{1, \ldots, j-1\} \), we note that after conditioning on \( \{x_{l,S_j}^{(l)} : l = 1, 2; i = 1, \ldots, n_1\} \), the only random variable in \( \hat{\epsilon}_j \) is \( \{\epsilon_{ij}, i = 1, \ldots, n_1+n_2\} \). Since \( \epsilon_{ij} \perp y_{il} \) for all \( l < j \) and \( i = 1, \ldots, n_1+n_2 \), we have

\[
\left( \tilde{\epsilon}_{j,S_j}, \hat{\epsilon}_j^{(1)} - \hat{\epsilon}_j^{(2)} \right) \big| A_{S_j} \]

\[
d \sim N \left( \tilde{\epsilon}_{j,S_j}, \hat{\epsilon}_j^{(1)} - \hat{\epsilon}_j^{(2)} \big| A_{S_j} \right)
\]

for any \( S_j \subseteq E \subseteq \{1, 2, \ldots, j-1\} \). This completes the proof.

(2a) By (6) and (8), we immediately know that

\[
\tilde{a}_{j,S_j} \big| A_E \sim N \left( \tilde{a}_{j,S_j}, r_j FF^\top \right).
\]

We next find the explicit form of \( FF^\top \). Note that

\[
FF^\top = \left( X_{S_j}^\top (I_{n_1+n_2} - \mathcal{P}_W X_{S_j}) \right)^{-1}.
\]
Also,

\[ P_W = W(W^TW)^{-1}W^T = \frac{1}{n_1n_2} W \begin{bmatrix} n_1 & -n_1 \\ -n_1 & n_1 + n_2 \end{bmatrix} W^T = \frac{1}{n_2} (1_{n_1+n_2} - g)(1_{n_1+n_2} - g)^T + \frac{1}{n_1} gg^T. \]

Note that

\[ \frac{1}{n_2} X_{S_j}^T (1_{n_1+n_2} - g) = x_{S_j}^{(2)}, \quad \frac{1}{n_1} X_{S_j}^T g = x_{S_j}^{(1)}. \]

Then, we have

\[ X_{S_j}^T (1_{n_1+n_2} - P_W) X_{S_j} = \sum_{l=1}^{m} \sum_{i=1}^{n} (x_{i,S_j}^{(l)} - x_{S_j}^{(l)})(x_{i,S_j}^{(l)} - x_{S_j}^{(l)})^T = (n_1 + n_2) \hat{\Sigma}_{S_j}. \]

Here, \( x_{S_j}^{(l)} \) and \( x_{i,S_j}^{(l)} \), respectively, denote the sub-vector of \( x^{(l)} \) and \( x_i^{(l)} \) with entries indexed by \( S_j \). This completes the proof.

(2b) Under Gaussianity, we only need to show that

\[ \text{Cov} \left( \hat{q}_{j,S_j}, x_j^{(l)} \mid A_E \right) = 0 \quad \text{for } l = 1, 2. \]

Using similar tricks to (1), we have

\[ \text{Cov} \left( \hat{q}_{j,S_j}, x_j^{(1)} \mid A_E \right) = r_j F g = 0, \quad \text{Cov} \left( \hat{q}_{j,S_j}, x_j^{(2)} \mid A_E \right) = r_j F (1_{n_1+n_2} - g) = 0. \]

This completes the proof.

(2c) Since \( Q \) is strictly upper triangular, thus \( j \notin S_k \) but possibly \( k \in S_j \). Then, conditioned on \( A_{S_j \cup S_k \cup \{k\}} \), \( \hat{q}_{k,S_k} \) becomes a fixed quantity; however, since \( k < j \) and \( S_j \subseteq S_j \cup S_k \cup \{k\} \subseteq \{1, \ldots, j-1\} \), it follows from 2(a) that \( \hat{q}_{j,S_j} \) follows a multivariate normal distribution with mean \( q_{j,S_j} \) and covariance \( r_j (n_1 + n_2)^{-1} \{\hat{\Sigma}_{S_j}^{-1}\}^{-1}. \)

9
Therefore,

\[
\mathbb{E} \left[ (\hat{q}_{j,S_j} - q_{j,S_j})(\hat{q}_{k,S_k} - q_{k,S_k})^\top \mid A_{S_j \cup S_k \cup \{k\}} \right] = 0 \times (\hat{q}_{k,S_k} - q_{k,S_k})^\top = 0,
\]

which completes the proof.

(3a) Using the properties of the OLS variance estimator, conditioned on \( A_E \), we have \((n_1 + n_2 - |S_j| - 4)\hat{r}_j^{-1} \sim \chi^2_{n_1 + n_2 - 2 - |S_j|}\). This indicates that conditioned on \( A_E \), \((n_1 + n_2 - |S_j| - 4)^{-1}\hat{r}_j^{-1} r_j \) follows an inverse chi-square distribution with the degree of freedom \(n_1 + n_2 - |S_j| - 2\). Then, we know that \( \mathbb{E} \left[ (n_1 + n_2 - |S_j| - 4)^{-1}\hat{r}_j^{-1} r_j \mid A_E \right] = (n_1 + n_2 - |S_j| - 2)(n_1 + n_2 - |S_j| - 6)^{-1}\).

This leads to

\[
\mathbb{E} \left( \hat{r}_j^{-1} \mid A_E \right) = r_j^{-1}, \quad \text{Var} \left( \hat{r}_j^{-1} \mid A_E \right) = r_j^{-2}(n_1 + n_2 - |S_j| - 6)^{-1},
\]

which completes the proof.

(3b) Using similar arguments to (2c), we have for \( k < j \),

\[
\mathbb{E} \left[ (\hat{r}_j^{-1} - r_j^{-1})(\hat{r}_k^{-1} - r_k^{-1}) \mid A_{S_j \cup S_k \cup \{k\}} \right] = \mathbb{E} \left[ \hat{r}_j^{-1} - r_j^{-1} \mid A_{S_j \cup S_k \cup \{k\}} \right] \times (\hat{r}_k^{-1} - r_k^{-1}) = 0,
\]

which completes the proof.

With all the lemmas, we next prove Theorem 1.

**Proof of Theorem 1.** Recall that \( N = n_1 n_2 / (n_1 + n_2) \) and \( A_E = \{ x_{i,E}^{(l)} : l = 1, 2; i = 1, \ldots, n_l \} \)
for any set $E \subset \{1, \ldots, p\}$. We write

$$T^2(\hat{Q}, \hat{R}) = N(\overline{x}^{(1)} - \overline{x}^{(2)})^\top \Sigma^{-1}(\overline{x}^{(1)} - \overline{x}^{(2)})$$

$$+ N(\overline{x}^{(1)} - \overline{x}^{(2)})^\top \left\{ \Sigma_{\text{DMG}}^{-1} - \Sigma^{-1} \right\} (\overline{x}^{(1)} - \overline{x}^{(2)})$$

$$= J_1 + J_2. \quad (9)$$

It is easy to check that $J_1 \overset{\text{d}}{\sim} \chi^2_p$. Next, we provide an asymptotic bound for $J_2$. We first write

$$J_2 = N \left\{ \sum_{j \in \mathcal{J}} \hat{r}_j^{-1} \{ (\overline{x}^{(1)} - \overline{x}^{(2)})^\top (\epsilon_j - \hat{q}_j) \}^2 - \sum_{j \in \mathcal{J}} r_j^{-1} \{ (\overline{x}^{(1)} - \overline{x}^{(2)})^\top (\epsilon_j - q_j) \}^2 \right\}$$

$$+ N \left\{ \sum_{j \notin \mathcal{J}} \hat{r}_j^{-1} \{ (\overline{x}^{(1)} - \overline{x}^{(2)})^\top \epsilon_j \}^2 - \sum_{j \notin \mathcal{J}} r_j^{-1} \{ (\overline{x}^{(1)} - \overline{x}^{(2)})^\top \epsilon_j \}^2 \right\}$$

$$= J_{21} + J_{22}. \quad (10)$$

For $J_{22}$, based on Lemma 1.3, we have

$$\mathbb{E}(J_{22}) = N \sum_{j \notin \mathcal{J}} \mathbb{E} \left[ \mathbb{E} \left[ \{ (\overline{x}^{(1)} - \overline{x}^{(2)})^\top \epsilon_j \}^2 \mid A_{S_j} \right] \mathbb{E}(\hat{r}_j^{-1} - r_j^{-1} \mid A_{S_j}) \right] = 0. \quad (11)$$

This double expectation trick will be used throughout the proof. Also, we have

$$\text{Var}(J_{22}) = N^2 \sum_{j \notin \mathcal{J}} \mathbb{E} \left[ \{ (\overline{x}^{(1)} - \overline{x}^{(2)})^\top \epsilon_j \hat{r}_j^{-1/2} \}^4 \hat{r}_j^2 (\hat{r}_j^{-1} - r_j^{-1})^2 \right]$$

$$+ 2N^2 \sum_{j \notin \mathcal{J}} \sum_{k \notin \mathcal{J}, k < j} \mathbb{E} \left[ \{ (\overline{x}^{(1)} - \overline{x}^{(2)})^\top \epsilon_j \hat{r}_j^{-1/2} \}^2 (\overline{x}^{(1)} - \overline{x}^{(2)})^\top \epsilon_k \hat{r}_k^{-1/2} \}^2 \times$$

$$r_j (\hat{r}_j^{-1} - r_j^{-1}) r_k (\hat{r}_k^{-1} - r_k^{-1}) \right]$$

$$= J_{221} + J_{222}. \quad (12)$$
Since $|S_j| = 0$ for $j \notin \mathcal{J}$, using Lemma 1.3 and the double expectation trick, we have

$$J_{221} = N^2 \sum_{j \notin \mathcal{J}} \mathbb{E}\left[ \{(x^{(1)}_j - x^{(2)}_j)^\top e_j r_j^{-1/2}\}^4 \right] \mathbb{E}\left[ r_j^2 (\hat{r}_j - r_j)^2 \right] = \frac{3(p - p_0)}{n_1 + n_2 - 6}.$$

Here we also use the fact that $\sqrt{N}(x^{(1)}_j - x^{(2)}_j)^\top e_j r_j^{-1/2} \sim N(0, 1)$ for $j \notin \mathcal{J}$. Similarly, we have

$$J_{222} = 2N^2 \sum_{j \notin \mathcal{J}} \sum_{k \notin \mathcal{J}, k < j} \mathbb{E}\left[ \mathbb{E}\left[ \{(x^{(1)}_j - x^{(2)}_j)^\top e_j r_j^{-1/2}\}^2 r_j \mid B_{j,k} \right] \mathbb{E}\left[ (\hat{r}_j - r_j)^2 \mid B_{j,k} \right] \right] \times \left\{ \{(x^{(1)}_j - x^{(2)}_j)^\top e_k r_k^{-1/2}\}^2 r_k (\hat{r}_k - r_k) \right\} = 0.

Combining results of $J_{221}$ and $J_{222}$, we have

$$\text{Var}(J_2) = 3(p - p_0)/(n_1 + n_2 - 6) = o(1).$$

For $J_{21}$, we first write that

$$J_{21} \leq 2N \left\{ \sum_{j \in \mathcal{J}} \hat{r}_j^{-1} \{(x^{(1)}_j - x^{(2)}_j)^\top (e_j - q_j)\}^2 - \sum_{j \in \mathcal{J}} r_j^{-1} \{(x^{(1)}_j - x^{(2)}_j)^\top (e_j - q_j)\}^2 \right\} + 2N \sum_{j \in \mathcal{J}} \hat{r}_j^{-1} \{(x^{(1)}_j - x^{(2)}_j)^\top (\hat{q}_j - q_j)\}^2 = J_{211} + J_{212}. \quad (13)$$

For $J_{211}$, using Lemma 1.3, we have

$$\mathbb{E}(J_{211}) = N \left\{ \sum_{j \in \mathcal{J}} \mathbb{E}\left[ \mathbb{E}[\hat{r}_j^{-1} - r_j^{-1} \mid A_{S_j}] \mathbb{E}[\{(x^{(1)}_j - x^{(2)}_j)^\top (e_j - q_j)\}^2 \mid A_{S_j}] \right] \right\} = 0.$$
Also,

\[
\text{Var}(J_{211}) = N^2 \sum_{j \in J} \mathbb{E} \left[ \{(\overline{x}^{(1)} - \overline{x}^{(2)})^\top (e_j - q_j) r_j^{-1/2}\}^4 r_j^2 (\hat{r}_j^{-1} - r_j^{-1})^2 \right] \\
+ 2N^2 \sum_{j \in J} \sum_{k \in J, k < j} \mathbb{E} \left[ \{(\overline{x}^{(1)} - \overline{x}^{(2)})^\top (e_j - q_j) r_j^{-1/2}\}^2 \{(\overline{x}^{(1)} - \overline{x}^{(2)})^\top (e_k - q_k) r_k^{-1/2}\}^2 \\
\times r_j (\hat{r}_j^{-1} - r_j^{-1}) r_k (\hat{r}_k^{-1} - r_k^{-1}) \right].
\]

Noting that \( \sqrt{N}(\overline{x}^{(1)} - \overline{x}^{(2)})^\top (e_j - q_j) r_j^{-1/2} \sim N(0, 1) \) for \( j \in J \), we use the same tricks as those for \( J_{221} \) and \( J_{222} \) to show

\[
\text{Var}(J_{211}) = \sum_{j \in J} 3/(n_1 + n_2 - |S_j| - 6) \leq 3p_0/(n_1 + n_2 - d - 6).
\]

For \( J_{212} \), we first note that \( (\overline{x}^{(1)} - \overline{x}^{(2)})^\top (\hat{q}_j - q_j) = (\overline{x}^{(1)} - \overline{x}^{(2)})^\top \{\hat{\Sigma}_{S_j, S_j}\}^{-1/2}\{\hat{\Sigma}_{S_j, S_j}\}^{1/2}(\hat{q}_j - q_j)_{S_j} \) for \( j \in J \). For ease of notation, we let \( \nu_j^x = (\overline{x}^{(1)} - \overline{x}^{(2)})^\top \{\hat{\Sigma}_{S_j, S_j}\}^{-1/2} \) and \( \theta_j = \{\hat{\Sigma}_{S_j, S_j}\}^{1/2}(\hat{q}_j - q_j)_{S_j} \). It follows from Lemma 1.3 that \( \mathbb{E}(\theta_j | A_{S_j}) = 0 \) and \( \text{Cov}(\theta_j | A_{S_j}) = r_j(n_1 + n_2)^{-1} I_{S_j} \). Thus,

\[
\mathbb{E}(J_{212}) = N \sum_{j \in J} \sum_{s \in S_j, t \in S_j} \mathbb{E} \left[ \nu_j^x \nu_j^y \theta_j \theta_j \right] \\
= N \sum_{j \in J} \sum_{s \in S_j, t \in S_j} \mathbb{E} \left[ \mathbb{E}[\hat{r}_j^{-1} \theta_j \theta_j | A_{S_j}] \nu_j^x \nu_j^y \right] \\
= N \sum_{j \in J} \sum_{s \in S_j} \mathbb{E} \left[ \mathbb{E}[\theta_j^2 | A_{S_j}] \nu_j^x \nu_j^y \right] \\
= \frac{N}{n_1 + n_2} \sum_{j \in J} \mathbb{E}[\nu_j^x \nu_j^y] \\
= \sum_{j \in J} \frac{|S_j|}{n_1 + n_2} \left( 1 + N \mathbb{E} \left[ (\overline{x}^{(1)} - \overline{x}^{(2)})^\top S_j^{-1/2} \{\hat{\Sigma}_{S_j, S_j}\}^{-1/2} \{\hat{\Sigma}_{S_j, S_j}\}^{1/2} - I_{S_j} \right] S_j^{-1/2} (\overline{x}^{(1)} - \overline{x}^{(2)})_{S_j} \right).
\]
Since $d \log p_0 = o(n_1 + n_2)$, using Lemma 1.2, we have

$$NE\left[\left(\bar{X}^{(1)} - \bar{X}^{(2)}\right)^T \sum_{S_j, S_j}^{-1/2} \sum_{S_j, S_j}^{1/2} \hat{S}_{S_j, S_j} \{ \hat{S}_{S_j, S_j} - I\hat{S}_{S_j}\}^{-1/2} \sum_{S_j, S_j}^{-1/2} (\bar{X}^{(1)} - \bar{X}^{(2)}) s_j\right]$$

$$\leq NE\left[\max_{j \in J} \left\{\sum_{S_j, S_j}^{1/2} \hat{S}_{S_j, S_j} \{ \hat{S}_{S_j, S_j} - I\hat{S}_{S_j}\}^{-1/2} \sum_{S_j, S_j}^{1/2} (\bar{X}^{(1)} - \bar{X}^{(2)}) s_j\right\}\right]$$

$$\leq O_p\left(\frac{d \log p_0}{n_1 + n_2}\right).$$

Since $d \log p_0(n_1 + n_2)^{-1} = O(1)$ and $N_e = \sum_{j \in J} |S_j|$, we have

$$\mathbb{E}(J_{212}) = O\left(\frac{N_e}{n_1 + n_2}\right).$$

Furthermore,

$$\mathbb{E}(J_{212}^2) = N^2 \mathbb{E}\left[\left(\sum_{j \in J, k \in J} \hat{r}_j^{-1}\{\{\bar{X}^{(1)} - \bar{X}^{(2)}\}^T(\hat{q}_j - q_j)\}\right)^2\right]$$

$$= N^2 \sum_{j \in J, k \in J} \mathbb{E}\left[\hat{r}_j^{-1}\{\sum_{s \in S_j, t \in S_j} \nu_{js \theta_js, \theta_jt} \times \hat{r}_k^{-1}\sum_{u \in S_k, v \in S_k} \nu_{ku \theta_ku, \theta_kv}\}ight]$$

$$= N^2 \sum_{j \in J} \mathbb{E}\left[\hat{r}_j^{-2}\{\sum_{s \in S_j, t \in S_j} \nu_{js \theta_js, \theta_jt}\}^2\right]$$

$$+ 2N^2 \sum_{j \in J, k \in J, k \neq j} \mathbb{E}\left[\hat{r}_j^{-1}\{\sum_{s \in S_j, t \in S_j} \nu_{js \theta_js, \theta_jt} \times \hat{r}_k^{-1}\sum_{u \in S_k, v \in S_k} \nu_{ku \theta_ku, \theta_kv}\}\right]$$

$$= J_{2121} + J_{2122}.$$
For $J_{2121}$, we have

\[
J_{2121} = N^2 \sum_{j \in J} \mathbb{E} \left[ \mathbb{E} \left( \tilde{r}_j^{-2} \mid A_{S_j} \right) \mathbb{E} \left( \left( \sum_{s \in S_j, t \in S_j} \nu_{s} \theta_{s} \nu_{t} \theta_{t} \right)^2 \mid A_{S_j} \right) \right]
\]

\[
= N^2 \sum_{j \in J} r_j^{-2} \left( 1 + \frac{1}{n_1 + n_2 - |S_j| - 6} \right) \sum_{s, t, a, \nu \in S_j} \mathbb{E} \left[ \nu_{s} \nu_{t} \nu_{a} \nu_{\nu} \mathbb{E} \left( \theta_{s} \theta_{t} \theta_{a} \theta_{\nu} \mid A_{S_j} \right) \right].
\]

Since given $A_{S_j}$, $\theta_s \sim N(0, r_j(n_1 + n_2)^{-1})$ and $\theta_s \perp \theta_t$ for all $s \neq t$ and $s, t \in S_j$,

\[
J_{2121} = N^2 \sum_{j \in J} \frac{1}{(n_1 + n_2)^2} \left( 1 + \frac{1}{n_1 + n_2 - |S_j| - 6} \right) \left( \sum_{s \in S_j} \mathbb{E} \left[ \nu_{s}^4 \right] + \sum_{s, t, a, \nu \in S_j, s \neq t} \mathbb{E} \left[ \nu_{s}^2 \nu_{t}^2 \right] \right).
\]

Since $\sqrt{N} \nu_s \sim N(0, 1)$ for all $s$, and $\nu_{s,t} \perp \nu_{s,t}$ for all $s \neq t$, we have

\[
J_{2121} = O \left( \frac{dN}{(n_1 + n_2)^2} \right).
\]

We can use similar tricks on $J_{2122}$. Specifically,

\[
J_{2122} = 2N^2 \sum_{j \in J, \nu \in J, \kappa < j} \mathbb{E} \left[ \mathbb{E}_{\tilde{r}_j^{-1}} \sum_{s \in S_j, t \in S_j} \nu_{s} \theta_{s} \nu_{t} \theta_{t} \mid B_{j,k} \right] \times \mathbb{E}_{\tilde{r}_k^{-1}} \sum_{u \in S_k, v \in S_k} \nu_{u} \theta_{u} \nu_{v} \theta_{v}
\]

\[
= 2N^2 \sum_{j \in J, \nu \in J, \kappa < j} \mathbb{E} \left[ \tilde{r}_j^{-1} \sum_{s \in S_j} \nu_{s}^2 \mathbb{E}_{\tilde{r}_j^{-1}} \left[ \theta_{s}^2 \mid B_{j,k} \right] \right] \times \mathbb{E}_{\tilde{r}_k^{-1}} \sum_{u \in S_k, v \in S_k} \nu_{u} \theta_{u} \nu_{v} \theta_{v}
\]

\[
= \frac{2N}{n_1 + n_2} \sum_{j \in J, \nu \in J, \kappa < j} |S_j| \mathbb{E} \left[ \tilde{r}_k^{-1} \sum_{u \in S_k, v \in S_k} \nu_{u} \theta_{u} \nu_{v} \theta_{v} \right]
\]

\[
= \frac{2N}{n_1 + n_2} \sum_{j \in J, \nu \in J, \kappa < j} |S_j| \mathbb{E} \left[ \tilde{r}_k^{-1} \sum_{u \in S_k, v \in S_k} \nu_{u} \theta_{u} \nu_{v} \theta_{v} \right]
\]

\[
= \frac{2}{(n_1 + n_2)^2} \sum_{j \in J, \nu \in J, \kappa < j} |S_j||S_k| \leq \frac{N^2}{(n_1 + n_2)^2}.
\]
Since $d \leq N_e$, we have

$$\text{Var}(J_{212}) \leq \mathbb{E}(J_{212}^2) = O\left(\frac{N_e^2}{(n_1 + n_2)^2}\right).$$

Combining all the assertions above, we have

$$\mathbb{E}(J_2) = O\left(\frac{N_e}{n_1 + n_2}\right), \quad \text{Var}(J_2) = O\left(\frac{p}{n_1 + n_2 - d - 6} + \frac{N_e^2}{(n_1 + n_2)^2}\right).$$

Therefore, as $n_1, n_2, p \to \infty$,

- if $N_e = o(n_1 + n_2)$ and $p = o(n_1 + n_2)$, then under $H_0$, $T^2(\hat{Q}, \hat{R}) \to \chi^2_p$ in distribution.

- if $N_e = o\left(\sqrt{p}(n_1 + n_2)\right)$, we have

$$\frac{T^2(\hat{Q}, \hat{R}) - p}{\sqrt{2p}} = \frac{J_1 - p}{\sqrt{2p}} + o_p(1) \to N(0, 1)$$

in distribution.

This completes the proof.

---

Proof of Theorem 2. For ease of notation, let $d_\mu = \mu_1 - \mu_2$ and $y_d = x^{(1)} - x^{(2)} - d_\mu$. Recall that $N = n_1n_2(n_1 + n_2)^{-1}$, $D_{KL} = \sum_{j=1}^p r_j^{-1}\{d_\mu^T(e_j - q_j)\}^2$ and $D_{DAG} = \sum_{j\in\mathcal{J}} d_{\mu, S_j}^T\{\Sigma S_j - S_j\}^{-1}d_{\mu, S_j}$.

We can then write

$$T^2(\hat{Q}, \hat{R}) = Ny_d^T\Sigma_{DAG}^{-1}y_d + ND_{KL} + Nd_\mu^T(\Sigma_{DAG}^{-1} - \Sigma^{-1})d_\mu + 2Nd_{\mu, S_j}^T\Sigma_{DAG}^{-1}y_d. \quad (14)$$
Using (14), we can write
\[
\frac{T^2(\hat{Q}, R) - p}{\sqrt{2p}} \geq ND_{KL} - \frac{N y_d^\top \hat{\Sigma}^{-1} y_d - p}{\sqrt{2p}} - \frac{2N}{\sqrt{2p}} d^\top_{\mu,S\hat{\Sigma}^{-1}S\hat{\Sigma}^{-1}} - \frac{N}{\sqrt{2p}} d^\top_{\mu,\hat{\Sigma}^{-1}S\hat{\Sigma}^{-1}} \Delta_{\hat{\Sigma}^{-1}S\hat{\Sigma}^{-1}} d_{\mu}.
\]
\[\text{(15)}\]

Since \(d \log p_0 = O(n_1 + n_2)\) and \(N_e = o(\sqrt{p(n_1 + n_2)})\), one can use the proof of Theorem 1 to show that \(
\frac{N y_d^\top \hat{\Sigma}^{-1} y_d - p}{\sqrt{2p}} \overset{d}{\to} N(0, 1)\) as \(n_1, n_2 \to \infty\).

We now show \(
\frac{N}{\sqrt{2p}} d^\top_{\mu,S\hat{\Sigma}^{-1}S\hat{\Sigma}^{-1}} d_{\mu} = o_p(1).\) Specifically,
\[
d^\top_{\mu,S\hat{\Sigma}^{-1}S\hat{\Sigma}^{-1}} d_{\mu} = \sum_{j=1}^p r_j^{-1} (d^\top_{\mu}(e_j - \hat{q}_j))^2 - \sum_{j=1}^p r_j^{-1} (d^\top_{\mu}(e_j - q_j))^2.
\]

Again, using the double expectation trick, we have
\[
E\left[\hat{r}_j^{-1} (d^\top_{\mu}(e_j - \hat{q}_j))^2\right] = E\left[E\left[\hat{r}_j^{-1} \mid A_{S_j}\right] E\left[\left(d^\top_{\mu}(e_j - \hat{q}_j)\right)^2 \mid A_{S_j}\right]\right]
\]
\[
= r_j^{-1} E\left[\left(d^\top_{\mu}(e_j - q_j)\right)^2 + r_j(n_1 + n_2)^{-1} d^\top_{\mu,S_j} (\hat{\Sigma}^{-1} S_{\hat{S}_j,S_j})^{-1} d_{\mu,S_j}\right].
\]

Since \(d \log p_0 = o(n_1 + n_2)\), using Lemma 1.2, we have
\[
E\left[d^\top_{\mu,S\hat{\Sigma}^{-1}S\hat{\Sigma}^{-1}} d_{\mu}\right] = \frac{1}{n_1 + n_2} \sum_{j \in \mathcal{J}} E\left[d^\top_{\mu,S_j} (\hat{\Sigma}^{-1} S_{\hat{S}_j,S_j})^{-1} d_{\mu,S_j}\right]
\]
\[
= D_{\text{DAG}} \frac{1}{n_1 + n_2} \times O\left(1 + \frac{d \log p_0}{(n_1 + n_2)}\right).
\]

Since \(D_{\text{DAG}} = O(p^{1/2} N^{-\gamma})\) for \(\gamma > 1/2\) and \(N/(n_1 + n_2) < 1\), we have
\[
\frac{N}{\sqrt{2p}} E\left[d^\top_{\mu,S\hat{\Sigma}^{-1}S\hat{\Sigma}^{-1}} d_{\mu}\right] = o(1).
\]

17
Also,

\[
\mathbb{E} \left[ \{d^T \mu, (\Sigma^{-1} - \Sigma)^{-1} d\mu\}^2 \right] \leq 2 \mathbb{E} \left( \sum_{j=1}^p (\hat{r}_j - r_j^{-1}) \{d^T u(e_j - \hat{q}_j)\}^2 \right)
\]

\[
+ 2 \mathbb{E} \left\{ \sum_{j=1}^p r_j^{-1} \left( \{d^T u(e_j - \hat{q}_j)\}^2 - \{d^T u(e_j - q_j)\}^2 \right) \right\}^2
\]

\[= 2(J_3 + J_4).\]

Note that

\[
J_3 = \sum_{j=1}^p \mathbb{E} \left[ (\hat{r}_j - r_j^{-1})^2 \{d^T u(e_j - \hat{q}_j)\}^4 \right]
\]

\[
+ 2 \sum_{k<j} \mathbb{E} \left[ (\hat{r}_j - r_j^{-1}) \{d^T u(e_j - \hat{q}_j)\}^2 (\hat{r}_k - r_k^{-1}) \{d^T u(e_k - \hat{q}_k)\}^2 \right]
\]

\[= J_{31} + 2J_{32}.\]

For \(J_{31}\), we have

\[
J_{31} = \sum_{j=1}^p \frac{r_j^{-2}}{n_1 + n_2 - |S_j| - 6} \mathbb{E} \left[ \{d^T u(e_j - \hat{q}_j)\}^4 \mid A_{S_j} \right]
\]

\[
\leq 8 \sum_{j=1}^p \frac{r_j^{-2}}{n_1 + n_2 - |S_j| - 6} \left( \{d^T u(e_j - q_j)\}^4 + 3 \mathbb{E} \left[ \left( \frac{r_j d^T u(\Sigma S_j \Sigma_j)^{-1} d \mu, S_j)}{n_1 + n_2} \right)^2 \right] \right) I(j \in \mathcal{J})
\]

\[
\leq \frac{8}{n_1 + n_2 - d - 6} \sum_{j=1}^p \frac{r_j^{-2}}{n_1 + n_2 - |S_j| - 6} \left( \{d^T u(e_j - q_j)\}^4 + 24 \sum_{j \in \mathcal{J}} \mathbb{E} \left[ \left( \frac{d^T \mu, S_j \{\Sigma S_j \Sigma_j \}^{-1} d \mu, S_j)}{n_1 + n_2} \right)^2 \right] \right). \]
Note that
\[
\sum_{j \in \mathcal{J}} \mathbb{E} \left[ \left( \frac{d_{\mu,Sj}}{n_1 + n_2} \right)^2 \right] \leq \frac{2}{(n_1 + n_2)^2} \sum_{j \in \mathcal{J}} \left( d_{\mu,Sj} \right)^2 \leq \frac{2}{(n_1 + n_2)^2} \sum_{j \in \mathcal{J}} \left( d_{\mu,Sj} \right)^2 \times O \left( 1 + \frac{d^2 \log^2 p_0}{(n_1 + n_2)^2} \right).
\]

Also,
\[
\sum_{j=1}^{p} r_j^{-2} (d_{\mu}^T (e_j - q_j))^4 \leq D_{KL}^2, \quad \sum_{j \in \mathcal{J}} \left( d_{\mu,Sj} \right)^2 \leq D_{DAG}^2.
\]

Since \( d \log p_0 = O(n_1 + n_2) \), we then have
\[
J_{31} = O \left( \frac{D_{KL}^2}{n_1 + n_2 - d - 6} + \frac{D_{DAG}^2}{(n_1 + n_2)^2(n_1 + n_2 - d - 6)} \right).
\]

Also, using the double expectation trick, it is easy to see that \( J_{32} = 0 \). Therefore,
\[
J_3 = O \left( \frac{D_{KL}^2}{n_1 + n_2 - d - 4} + \frac{D_{DAG}^2}{(n_1 + n_2)^2(n_1 + n_2 - d - 6)} \right).
\]

For \( J_4 \), we have
\[
J_4 = \sum_{j=1}^{p} \mathbb{E} \left[ r_j^{-2} \left( \left\{ \{d_u^T (e_j - \hat{q}_j)\}^2 - \{d_u^T (e_j - q_j)\}^2 \right\}^2 \right) \right]
+ 2 \sum_{k < j} \mathbb{E} \left[ r_j^{-1} \left( \left\{ \{d_u^T (e_j - \hat{q}_j)\}^2 - \{d_u^T (e_j - q_j)\}^2 \right\} r_k^{-1} \left( \left\{ \{d_u^T (e_k - \hat{q}_k)\}^2 - \{d_u^T (e_k - q_k)\}^2 \right\} \right) \right]
= J_{41} + 2J_{42}.
\]
For $J_{41}$, using similar tricks, we can show that

$$J_{41} \leq \frac{8}{n_1 + n_2} \sum_{j \in J} \frac{1}{r_j} \left\{ \mathbf{d}_u^\top (\mathbf{e}_j - \mathbf{q}_j) \right\}^2 \mathbf{d}_{\mu,S_j}^\top \left\{ \Sigma_{S_j,S_j} \right\}^{-1} \mathbf{d}_{\mu,S_j} \times O \left( 1 + \frac{d \log p_0}{n_1 + n_2} \right)$$

$$+ \frac{6}{(n_1 + n_2)^2} \sum_{j \in J} \left( \mathbf{d}_{\mu,S_j}^\top \left\{ \Sigma_{S_j,S_j} \right\}^{-1} \mathbf{d}_{\mu,S_j} \right)^2 \times O \left( 1 + \left( \frac{d \log p_0}{n_1 + n_2} \right)^2 \right)$$

$$= O \left( \frac{D_{DL,DAG}}{n_1 + n_2} + \frac{D_{DAG}^2}{(n_1 + n_2)^2} \right).$$

and

$$2J_{42} = \frac{1}{(n_1 + n_2)^2} \sum_{k \neq j} \mathbf{d}_{\mu,S_j}^\top \left\{ \Sigma_{S_j,S_j} \right\}^{-1} \mathbf{d}_{\mu,S_j} \mathbf{d}_{\mu,S_k}^\top \left\{ \Sigma_{S_k,S_k} \right\}^{-1} \mathbf{d}_{\mu,S_k} \times O \left( 1 + \frac{d \log p_0}{n_1 + n_2} \right)^2$$

$$= O \left( \frac{D_{DAG}^2}{(n_1 + n_2)^2} \right).$$

Thus, since $D_{DAG} \approx p^\beta$, $D_{KL} \approx p^{1/2} N^{-\gamma}$ and $d = o(n_1 + n_2)$ for $\gamma > 1/2$ and $0 < \beta < 1/2$, combining the results for $J_3$ and $J_4$, one can see that

$$\frac{N^2}{2p} \mathbb{E} \left[ \left\{ \mathbf{d}_{\mu}^\top (\hat{\Sigma}_{DAG}^{-1} - \Sigma^{-1}) \mathbf{d}_\mu \right\}^2 \right] = o(1).$$

Therefore,

$$\frac{N}{\sqrt{2p}} \left( \mathbf{d}_{\mu}^\top (\hat{\Sigma}_{DAG}^{-1} - \Sigma^{-1}) \mathbf{d}_\mu \right) = o_p(1).$$

We now focus on $\mathbf{d}_{\mu}^\top \hat{\Sigma}_{DAG}^{-1} \mathbf{y}_d$. Similarly, we can write

$$\mathbf{d}_{\mu}^\top \hat{\Sigma}_{DAG}^{-1} \mathbf{y}_d = \mathbf{d}_{\mu}^\top \Sigma^{-1} \mathbf{y}_d + \mathbf{d}_{\mu}^\top (\hat{\Sigma}_{DAG}^{-1} - \Sigma^{-1}) \mathbf{y}_d = J_5 + J_6$$

For $J_5$, it is easy to see that $J_5 \sim N(0, N^{-1} D_{KL})$, this indicates that $N J_5 / \sqrt{2p} = o_p(1)$, since
\[ D_{\text{KL}} = O(p^{1/2}N^{-\gamma}) \] for \( \gamma > 1/2 \). For \( J_6 \), we first write
\[
J_6 = \sum_{j=1}^{p} \hat{r}_j^{-1} d^\top_\mu (e_j - \hat{q}_j) (e_j - \hat{q}_j)^\top y_d - \sum_{j=1}^{p} r_j^{-1} d^\top_\mu (e_j - q_j) (e_j - q_j)^\top y_d.
\]

Again, using the double expectation trick and Lemmas 1.2 and 1.3, we have
\[
\mathbb{E}[J_6] = \sum_{j=1}^{p} \mathbb{E} \left[ (e_j - \hat{q}_j)(e_j - \hat{q}_j)^\top y_d \mid A_{S_j} \right] d^\top_\mu \mathbb{E} \left[ (e_j - \hat{q}_j)(e_j - \hat{q}_j)^\top y_d \mid A_{S_j} \right] \\
= -2 \sum_{j=1}^{p} r_j^{-1} \mathbb{E} \left[ d^\top_\mu (e_j - q_j)(\hat{q}_j - q_j)^\top y_d \right] + \sum_{j=1}^{p} r_j^{-1} \mathbb{E} \left[ d^\top_\mu (\hat{q}_j - q_j)(\hat{q}_j - q_j)^\top y_d \right] \\
= \frac{1}{n_1 + n_2} \sum_{j \in J} \mathbb{E} \left[ d^\top_{\mu, S_j} (\hat{\Sigma}_{S_j, S_j})^{-1} y_d,S_j \right] \\
= \frac{1}{(n_1 + n_2)N} \sum_{j \in J} d^\top_{\mu, S_j} (\hat{\Sigma}_{S_j, S_j})^{-1} d_{\mu, S_j} \times O \left( 1 + \frac{d \log p_0}{n_1 + n_2} \right) \\
= O \left( \frac{D_{\text{DAG}}}{(n_1 + n_2)N} \right).
\]

Since \( D_{\text{DAG}} = O(p^\beta) \) for \( 0 < \beta < 1/2 \), one can see that \( \sqrt{N} \mathbb{E}(J_6)/\sqrt{p} = o(1) \). Similarly,
\[
\mathbb{E}[J^2_6] \leq 2 \mathbb{E} \left[ \sum_{j=1}^{p} (\hat{r}_j^{-1} - r_j^{-1}) d^\top_\mu (e_j - q_j)(e_j - q_j)^\top y_d \right]^2 \\
+ 16 \mathbb{E} \left[ \sum_{j=1}^{p} \hat{r}_j^{-1} d^\top_\mu (e_j - q_j)(\hat{q}_j - q_j)^\top y_d \right]^2 \\
+ 4 \mathbb{E} \left[ \sum_{j=1}^{p} \hat{r}_j^{-1} d^\top_\mu (\hat{q}_j - q_j)(\hat{q}_j - q_j)^\top y_d \right]^2. \\
= 2J_{61} + 16J_{62} + 4J_{63}.
\]
For $J_{61}$, we have

$$J_{61} \leq \frac{1}{N(n_1 + n_2 - d - 6)} \sum_{j=1}^{P} r_j^{-2} \{d_j^T(e_j - q_j)\}^2 (e_j - q_j)^T \Sigma (e_j - q_j) \leq \frac{d_{KL}}{N(n_1 + n_2 - d - 6)}.$$  

Here, we use the fact that $(e_j - q_j)^T \Sigma (e_j - q_j) = r_j$. Using similar tricks, we have

$$J_{62} \leq \frac{D_{KL} d}{(n_1 + n_2) N} \times O \left(1 + \frac{d \log p_0}{n_1 + n_2}\right).$$

and

$$J_{63} = \sum_{j=1}^{P} r_j^{-2} (1 + \frac{1}{n_1 + n_2 - |S_j| - 6}) \mathbb{E} \left[ \{d_j^T(\hat{q}_j - q_j)(\hat{q}_j - q_j)^T y_d\}^2 \right]$$
$$+ \frac{2}{(n_1 + n_2)^2} \sum_{k<j} \mathbb{E} \left[ d_{\mu,S_j}^T \{\hat{\Sigma}_{S_j,S_j}\}^{-1} y_{d,S_j} d_{\mu,S_k}^T \{\hat{\Sigma}_{S_k,S_k}\}^{-1} y_{d,S_k} \right]$$
$$= J_{631} + J_{632}.$$  

For $J_{631}$, recall that $\theta_{S_j} = \{\hat{\Sigma}_{S_j,S_j}\}^{1/2}(\hat{q}_j - q_j)_{S_j}$. For ease of notations, let $\psi_j = \{\hat{\Sigma}_{S_j,S_j}\}^{-1/2} d_{\mu,S_j}$ and $\phi_j = \{\hat{\Sigma}_{S_j,S_j}\}^{-1/2} y_{d,S_j}$. Then, using the fact that $\theta_{S_j} | A_{S_j} \sim N(0, r_j (n_1 + n_2)^{-1})$, we have

$$\mathbb{E} \left[ \{d_j^T(\hat{q}_j - q_j)(\hat{q}_j - q_j)^T y_d\}^2 \right] = \mathbb{E} \left[ \sum_{s,t,u,v \in S_j} \phi_{js} \phi_{jt} \psi_{ju} \psi_{jv} \theta_{js} \theta_{jt} \theta_{ju} \theta_{jv} \right]$$
$$\leq \frac{4r_j^2}{(n_1 + n_2)^2} \mathbb{E} \left[ d_{\mu,S_j}^T \{\hat{\Sigma}_{S_j,S_j}\}^{-1} d_{\mu,S_j} \times y_{d,S_j}^T \{\hat{\Sigma}_{S_j,S_j}\}^{-1} y_{d,S_j} \right]$$
$$+ \frac{2r_j^2}{(n_1 + n_2)^2} \mathbb{E} \left[ (d_{\mu,S_j}^T \{\hat{\Sigma}_{S_j,S_j}\}^{-1} y_{d,S_j})^2 \right].$$
Since \( y_{d,S_j} \sim N(0, N^{-1} \Sigma_{S_j,S_j}) \), by some algebra, one can see that

\[
J_{631} \leq \frac{4d}{N(n_1+n_2)^2} \left( 1 + \frac{1}{n_1+n_2 - d - 4} \right) \sum_{j \in J} d^T_{\mu,S_j} \{ \Sigma_{S_j,S_j} \}^{-1} d_{\mu,S_j} \times O \left( \frac{1 + \log p_0}{n_1+n_2} \right)
\]
\[
+ \frac{2}{N(n_1+n_2)^2} \sum_{j \in J} d^T_{\mu,S_j} \{ \Sigma_{S_j,S_j} \}^{-1} d_{\mu,S_j} \times O \left( \frac{1 + \log p_0}{n_1+n_2} \right)
\]
\[
= O \left( \frac{4dD_{DAG}}{N(n_1+n_2)^2} \right).
\]

Similarly,

\[
J_{632} \leq \frac{1}{N(n_1+n_2)^2} \sum_{j \in J} d^T_{\mu,S_j} \{ \Sigma_{S_j,S_j} \}^{-1} d_{\mu,S_j} \times O \left( \frac{1 + \log p_0}{n_1+n_2} \right)
\]
\[
= \frac{D_{DAG}}{N(n_1+n_2)^2}.
\]

Since \( D_{KL} = O(p^{1/2}N^{-\gamma}) \), \( D_{DAG} = O(p^\beta) \) and \( d = o(n_1+n_2) \) for \( \gamma > 1/2 \) and \( 0 < \beta < 1/2 \), one can see that \( N^2 \mathbb{E}(J_6^2)/p = o(1) \). Combining all the assertions above, as \( n_1,n_2 \to \infty \), we have

\[
\mathbb{P} \left( \left| \frac{T^2(\hat{Q}, \hat{P}) - p}{\sqrt{2p}} \right| \geq z_{\alpha/2} \right) \geq \mathbb{P} \left( \left| Z \right| \leq \frac{ND_{KL}}{\sqrt{2p}} - z_{\alpha/2} \right).
\]

Furthermore, if \( 1/2 < \gamma < 1 \), then

\[
\mathbb{P} \left( \left| Z \right| \leq \frac{ND_{KL}}{\sqrt{2p}} - z_{\alpha/2} \right) = 1 \text{ as } n_1,n_2 \to \infty.
\]

This completes the proof. \( \square \)
2  More information on the simulation studies

2.1  Additional simulation results assuming accurate edge information

Figure S1. Empirical type I error rates ($\delta = 0$) and powers ($|\delta| > 0$) under $n_1 = n_2 = 50$ and $p = 20$, with the number of children nodes in the DAG set to $p_0 = 0.4p$ (columns 1 and 2) or $p_0 = 0.8p$ (columns 3 and 4), and the number of non-zero signals set to $q = 0.1p$ (columns 1 and 3) or $0.5p$ (columns 2 and 4).

Figure S2. Empirical type I error rates ($\delta = 0$) and powers ($|\delta| > 0$) under $n_1 = n_2 = 100$ and $p = 100$, with $p_0 = 0.4p$ (columns 1 and 2) or $p_0 = 0.8p$ (columns 3 and 4), and $q = 0.1p$ (columns 1 and 3) or $0.5p$ (columns 2 and 4).
Figure S3. Empirical type I error rates ($\delta = 0$) and powers ($|\delta| > 0$) under $n_1 = n_2 = 100$ and $p = 300$, with $p_0 = 0.4p$ (columns 1 and 2) or $p_0 = 0.8p$ (columns 3 and 4), and $q = 0.1p$ (columns 1 and 3) or $0.5p$ (columns 2 and 4).

2.2 Additional simulation results given mis-specified edge information

Figure S4. Empirical type I error rates ($\delta = 0$) and powers ($|\delta| > 0$) of T2-DAG $Z$ (top panel) and T2-DAG $\chi^2$ (bottom panel) under $n_1 = n_2 = 50$ and $p = 50$, with $p_0 = 0.8p$, and $q = 0.1p$ (left panel) or $0.5p$ (right panel). The T2-DAG was applied based on $A$ (“T2-DAG Correct”), $A_1$ (“T2-DAG Missing”), or $A_2$ (“T2-DAG Redundant”). The Graph T2 was applied based on the true adjacency matrix $A$. 

25
Figure S5. Empirical type I error rates ($\delta = 0$) and powers ($|\delta| > 0$) of T2-DAG $Z$ (top panel) and T2-DAG $\chi^2$ (bottom panel) under $n_1 = n_2 = 100$ and $p = 100$, with $p_0 = 0.8p$, and $q = 0.1p$ (left panel) or $0.5p$ (right panel). The T2-DAG was applied based on $A$ (“T2-DAG Correct”), $A_1$ (“T2-DAG Missing”), or $A_2$ (“T2-DAG Redundant”). The Graph T2 was applied based on the true adjacency matrix $A$. 

---

26
Figure S6. Empirical type I error rates ($\delta = 0$) and powers ($|\delta| > 0$) of T2-DAG $Z$ (top panel) and T2-DAG $\chi^2$ (bottom panel) under $n_1 = n_2 = 100$ and $p = 300$, with $p_0 = 0.8p$, and $q = 0.1p$ (left panel) or $0.5p$ (right panel). The T2-DAG was applied based on $A$ ("T2-DAG Correct"), $A_1$ ("T2-DAG Missing"), or $A_2$ ("T2-DAG Redundant"). The Graph T2 was applied based on the true adjacency matrix $A$. 

27
2.3 Additional simulation results in the presence of unadjusted confounding effects

Figure S7. Empirical type I error rates ($\delta = 0$) and powers ($|\delta| > 0$) of the various tests under $n_1 = n_2 = 50$ and $p = 50$ (column 1) or $n_1 = n_2 = 100$ and $p = 100$ (column 2), assuming two continuous confounders, with $p_0 = 0.8p$, and $q = 0.5p$.

2.4 Additional simulation results given non-Gaussian error terms

We conducted a sensitivity analysis on T2-DAG in terms of the distribution of the error terms. Specifically, we considered three alternative non-Gaussian distributions: (1) uniform distribution, where $\epsilon_{ij}$ was generated from $\text{Unif}(-\sqrt{3r_j}, \sqrt{3r_j})$, with a zero mean and variance equal to $r_j$ as in the simulation setting in Section 3.2 of the main manuscript; (2) Gamma distribution, where $\epsilon_{ij}$ was generated from a Gamma distribution with a shape parameter $s = 10$ and a scale parameter $\sqrt{r_j/s}$ and then subtracted by its mean, $s\sqrt{r_j/s}$, which gave a zero mean and variance equal to $r_j$; and (3) log-normal distribution, where $\epsilon_{ij}$ was generated from a log-normal distribution with zero mean and variance $\tau^2 = 0.16$ on the log scale and then subtracted by its mean, $\exp(\tau^2/2)$. 

28
Figure S8. Empirical type I error rates ($\delta = 0$) and powers ($|\delta| > 0$) of the various tests under $n_1 = n_2 = 50$ and $p = 40$, with error terms generated from log-normal distributions based on the raw data (row 1) or log-transformed data (row 2). The number of children nodes in the DAG is set to $p_0 = 0.6p$ (columns 1 and 2) or $p_0 = 0.8p$ (columns 3 and 4), and the number of non-zero signals is set to $q = 0.1p$ (columns 1 and 3) or $0.5p$ (columns 2 and 4).

Figure S9. Empirical type I error rates ($\delta = 0$) and powers ($|\delta| > 0$) of the various tests under $n_1 = n_2 = 100$ and $p = 100$, with error terms generated from log-normal distributions based on the raw data (row 1) or log-transformed data (row 2). The number of children nodes in the DAG is set to $p_0 = 0.6p$ (columns 1 and 2) or $p_0 = 0.8p$ (columns 3 and 4), and the number of non-zero signals is set to $q = 0.1p$ (columns 1 and 3) or $0.5p$ (columns 2 and 4).
Figure S10. Empirical type I error rates ($\delta = 0$) and powers ($|\delta| > 0$) of the various tests under $n_1 = n_2 = 100$ and $p = 300$, with error terms generated from log-normal distributions based on the raw data (row 1) or log-transformed data (row 2). The number of children nodes in the DAG is set to $p_0 = 0.6p$ (columns 1 and 2) or $p_0 = 0.8p$ (columns 3 and 4), and the number of non-zero signals is set to $q = 0.1p$ (columns 1 and 3) or $0.5p$ (columns 2 and 4).
Figure S11. Empirical type I error rates ($\delta = 0$) and powers ($|\delta| > 0$) of the various tests under $n_1 = n_2 = 50$ and $p = 40$, with error terms generated from gamma distributions based on the raw data (row 1) or log-transformed data (row 2). The number of children nodes in the DAG is set to $p_0 = 0.6p$ (columns 1 and 2) or $p_0 = 0.8p$ (columns 3 and 4), and the number of non-zero signals is set to $q = 0.1p$ (columns 1 and 3) or $0.5p$ (columns 2 and 4).

Figure S12. Empirical type I error rates ($\delta = 0$) and powers ($|\delta| > 0$) of the various tests under $n_1 = n_2 = 50$ and $p = 100$, with error terms generated from gamma distributions based on the raw data (row 1) or log-transformed data (row 2). The number of children nodes in the DAG is set to $p_0 = 0.6p$ (columns 1 and 2) or $p_0 = 0.8p$ (columns 3 and 4), and the number of non-zero signals is set to $q = 0.1p$ (columns 1 and 3) or $0.5p$ (columns 2 and 4).
Figure S13. Empirical type I error rates ($\delta = 0$) and powers ($|\delta| > 0$) of the various tests under $n_1 = n_2 = 100$ and $p = 100$, with error terms generated from gamma distributions based on the raw data (row 1) or log-transformed data (row 2). The number of children nodes in the DAG is set to $p_0 = 0.6p$ (columns 1 and 2) or $p_0 = 0.8p$ (columns 3 and 4), and the number of non-zero signals is set to $q = 0.1p$ (columns 1 and 3) or $0.5p$ (columns 2 and 4).
Figure S14. Empirical type I error rates ($\delta = 0$) and powers ($|\delta| > 0$) of the various tests under $n_1 = n_2 = 100$ and $p = 300$, with error terms generated from gamma distributions based on the raw data (row 1) or log-transformed data (row 2). The number of children nodes in the DAG is set to $p_0 = 0.6p$ (columns 1 and 2) or $p_0 = 0.8p$ (columns 3 and 4), and the number of non-zero signals is set to $q = 0.1p$ (columns 1 and 3) or $0.5p$ (columns 2 and 4).

Figure S15. Empirical type I error rates ($\delta = 0$) and powers ($|\delta| > 0$) of the various tests under $n_1 = n_2 = 50$ and $p = 40$, with error terms generated from uniform distributions. The number of children nodes in the DAG is set to $p_0 = 0.6p$ (columns 1 and 2) or $p_0 = 0.8p$ (columns 3 and 4), and the number of non-zero signals is set to $q = 0.1p$ (columns 1 and 3) or $0.5p$ (columns 2 and 4).
Figure S16. Empirical type I error rates ($\delta = 0$) and powers ($|\delta| > 0$) of the various tests under $n_1 = n_2 = 50$ and $p = 100$, with error terms generated from uniform distributions. The number of children nodes in the DAG is set to $p_0 = 0.6p$ (columns 1 and 2) or $p_0 = 0.8p$ (columns 3 and 4), and the number of non-zero signals is set to $q = 0.1p$ (columns 1 and 3) or $0.5p$ (columns 2 and 4).

Figure S17. Empirical type I error rates ($\delta = 0$) and powers ($|\delta| > 0$) of the various tests under $n_1 = n_2 = 100$ and $p = 100$, with error terms generated from uniform distributions. The number of children nodes in the DAG is set to $p_0 = 0.6p$ (columns 1 and 2) or $p_0 = 0.8p$ (columns 3 and 4), and the number of non-zero signals is set to $q = 0.1p$ (columns 1 and 3) or $0.5p$ (columns 2 and 4).
Figure S18. Empirical type I error rates ($\delta = 0$) and powers ($|\delta| > 0$) of the various tests under $n_1 = n_2 = 100$ and $p = 300$, with error terms generated from uniform distributions. The number of children nodes in the DAG is set to $p_0 = 0.6p$ (columns 1 and 2) or $p_0 = 0.8p$ (columns 3 and 4), and the number of non-zero signals is set to $q = 0.1p$ (columns 1 and 3) or $0.5p$ (columns 2 and 4).

3 Additional information on real data analysis

Figure S16: $-\log_{10}(p)$ of the various tests for 20 randomly selected pathways between stage I and stage II lung cancer.
Figure S17: The $-\log_{10}(p)$ of the various tests for 20 randomly selected pathways between stage I and stage IV lung cancer.

Figure S18: The $-\log_{10}(p)$ of the various tests for 20 randomly selected pathways between stage II and stage III lung cancer.
Figure S19: The $-\log_{10}(p)$ of the various tests for 20 randomly selected pathways between stage II and stage IV lung cancer.

Figure S20: The $-\log_{10}(p)$ of the various tests for 20 randomly selected pathways between stage III and stage IV lung cancer.
Figure S21: The $-\log_{10}(p)$ of the comparison between stage II and stage IV lung cancer ($N_2 = 124$, $N_4 = 24$) for the pathways identified by any test to have significantly different expression levels between the two cancer stages.

Figure S22: The $-\log_{10}(p)$ of the comparison between stage III and stage IV lung cancer ($N_3 = 84$, $N_4 = 27$) for the pathways identified by any test to have significantly different expression levels between the two cancer stages.

Detailed information on the pathways and testing results of data analysis are summarized in Supplementary Tables S1 - S10 which are provided in a separate file named “Supplemental Tables S1 - S10”.
Table S11: Results for testing differential expression of 206 pathways between stage I and stage II lung cancer ($N_1 = 278$, $N_2 = 124$). In each panel, the numbers in the diagonal cells are the numbers of significant pathways identified by the corresponding methods; the numbers in the $(i, j)$-th cell in the upper and lower triangle respectively indicate the number of significant pathways detected by both method $i$ and method $j$, and the correlation in $-\log_{10}(p)$ between the two methods.

|          | T2-DAG $\chi^2$ | T2-DAG $z$ | Graph T2 | T2 | CH-Q | SK | CLX | GCT | aSPU | ARHT |
|----------|-----------------|------------|----------|----|------|----|-----|-----|------|------|
| T2-DAG $\chi^2$ | 39             | 39         | 0        | 0  | 11   | 4  | 0   | 10  | 13   | 3    |
| T2-DAG $z$    | 0.98           | 57         | 0        | 0  | 13   | 4  | 0   | 10  | 18   | 3    |
| Graph T2      | 0.43           | 0.39       | 1        | 1  | 0    | 0  | 0   | 0   | 0    | 0    |
| T2            | 0.39           | 0.34       | 0.70     | 1  | 0    | 0  | 0   | 0   | 0    | 0    |
| CH-Q          | 0.71           | 0.71       | 0.30     | 0.34 | 16  | 4  | 0   | 6   | 13   | 3    |
| SK            | 0.72           | 0.71       | 0.34     | 0.43 | 0.83 | 4  | 0   | 3   | 4    | 2    |
| CLX           | 0.40           | 0.34       | 0.23     | 0.35 | 0.40 | 0.48 | 0   | 0   | 0    | 0    |
| GCT           | 0.86           | 0.80       | 0.37     | 0.38 | 0.60 | 0.65 | 0.26 | 10  | 6    | 3    |
| aSPU          | 0.28           | 0.30       | 0.21     | 0.22 | 0.60 | 0.42 | 0.35 | 0.13 | 26   | 3    |
| ARHT          | 0.63           | 0.57       | 0.56     | 0.84 | 0.71 | 0.69 | 0.46 | 0.59 | 0.44 | 3    |

Table S12: Results for testing potential differential expression of 206 pathways between stage I and stage IV lung cancer ($N_1 = 278$, $N_4 = 27$). In each panel, the numbers in the diagonal cells are the numbers of significant pathways identified by the corresponding methods; the numbers in the $(i, j)$-th cell in the upper and lower triangle respectively indicate the number of significant pathways detected by both method $i$ and method $j$, and the correlation in $-\log_{10}(p)$ between the two methods.

|          | T2-DAG $\chi^2$ | T2-DAG $z$ | Graph T2 | T2 | CH-Q | SK | CLX | GCT | aSPU | ARHT |
|----------|-----------------|------------|----------|----|------|----|-----|-----|------|------|
| T2-DAG $\chi^2$ | 55             | 55         | 4        | 9  | 4    | 4  | 1   | 16  | 16   | 8    |
| T2-DAG $z$    | 0.98           | 74         | 4        | 9  | 4    | 4  | 2   | 16  | 16   | 8    |
| Graph T2      | 0.54           | 0.55       | 4        | 2  | 3    | 2  | 0   | 2   | 3    | 3    |
| T2            | 0.43           | 0.41       | 0.49     | 12  | 2    | 2  | 0   | 0   | 3    | 7    |
| CH-Q          | 0.44           | 0.54       | 0.43     | 0.32 | 5   | 2  | 0   | 1   | 5    | 3    |
| SK            | 0.72           | 0.74       | 0.60     | 0.59 | 0.59 | 4  | 1   | 0   | 3    | 2    |
| CLX           | 0.36           | 0.36       | 0.26     | 0.31 | 0.22 | 0.57 | 3   | 0   | 0    | 0    |
| GCT           | 0.85           | 0.82       | 0.51     | 0.34 | 0.31 | 0.62 | 0.14 | 16  | 6    | 2    |
| aSPU          | 0.39           | 0.45       | 0.47     | 0.27 | 0.69 | 0.52 | 0.33 | 0.33 | 19   | 5    |
| ARHT          | 0.39           | 0.48       | 0.40     | 0.47 | 0.95 | 0.55 | 0.22 | 0.23 | 0.54 | 9    |
Table S13: Results for testing potential differential expression of 206 pathways between stage II and stage IV lung cancer ($N_2 = 124$, $N_4 = 27$). In each panel, the numbers in the diagonal cells are the numbers of significant pathways identified by the corresponding methods; the numbers in the $(i, j)$-th cell in the upper and lower triangle respectively indicate the number of significant pathways detected by both method $i$ and method $j$, and the correlation in $-\log_{10}(p)$ between the two methods.

|          | T2-DAG $\chi^2$ | T2-DAG $z$ | Graph T2 | T2 | CH-Q | SK | CLX | GCT | aSPU | ARHT |
|----------|-----------------|------------|----------|----|------|----|-----|-----|------|------|
| T2-DAG   |                  |            |          |    |      |    |     |     |      |      |
| $\chi^2$| 6               | 6          | 0        | 1  | 0    | 0  | 0   | 1   | 1    | 1    |
| $z$      | 0.97            | 13         | 0        | 1  | 0    | 0  | 0   | 1   | 3    | 1    |
| Graph T2 | 0.38            | 0.34       | 0        | 0  | 0    | 0  | 0   | 0   | 0    | 0    |
| T2       | 0.45            | 0.40       | 0.49     | 1  | 0    | 0  | 0   | 0   | 0    | 1    |
| CH-Q     | 0.60            | 0.58       | 0.36     | 0.30| 0    | 0  | 0   | 0   | 0    | 0    |
| SK       | 0.74            | 0.77       | 0.45     | 0.45| 0.55 | 0  | 0   | 0   | 0    | 0    |
| CLX      | 0.29            | 0.25       | 0.13     | 0.27| 0.19 | 0.44| 0   | 0   | 0    | 0    |
| GCT      | 0.58            | 0.67       | 0.20     | 0.03| 0.40 | 0.50| -0.11| 2   | 0    | 0    |
| aSPU     | 0.43            | 0.45       | 0.16     | 0.18| 0.64 | 0.47| 0.38| 0.29| 6    | 0    |
| ARHT     | 0.50            | 0.44       | 0.40     | 0.91| 0.42 | 0.51| 0.27| 0.11| 0.22 | 1    |

Table S14: Results for testing potential differential expression of 206 pathways between stage III and stage IV lung cancer ($N_3 = 84$, $N_4 = 24$). In each panel, the numbers in the diagonal cells are the numbers of significant pathways identified by the corresponding methods; the numbers in the $(i, j)$-th cell in the upper and lower triangle respectively indicate the number of significant pathways detected by both method $i$ and method $j$, and the correlation in $-\log_{10}(p)$ between the two methods.

|          | T2-DAG $\chi^2$ | T2-DAG $z$ | Graph T2 | T2 | CH-Q | SK | CLX | GCT | aSPU | ARHT |
|----------|-----------------|------------|----------|----|------|----|-----|-----|------|------|
| T2-DAG   |                  |            |          |    |      |    |     |     |      |      |
| $\chi^2$| 0               | 0          | 0        | 0  | 0    | 0  | 0   | 0   | 0    | 0    |
| $z$      | 0.55            | 0.01       | 0        | 0  | 0    | 0  | 0   | 0   | 0    | 0    |
| Graph T2 | 0.33            | 0.01       | 0        | 0  | 0    | 0  | 0   | 0   | 0    | 0    |
| T2       | 0.35            | -0.01      | 0.34     | 0  | 0    | 0  | 0   | 0   | 0    | 0    |
| CH-Q     | 0.50            | 0.13       | 0.35     | 0.22| 0    | 0  | 0   | 0   | 0    | 0    |
| SK       | 0.37            | 0.89       | -0.01    | -0.07| 0.12 | 0  | 0   | 0   | 0    | 0    |
| CLX      | 0.45            | 0.10       | 0.08     | 0.30| 0.22 | 0.07| 0   | 0   | 0    | 0    |
| GCT      | -0.21           | 0.53       | -0.20    | -0.26| -0.21 | 0.62| -0.27| 5   | 0    | 0    |
| aSPU     | 0.38            | 0.11       | 0.12     | 0.15| 0.47 | 0.05| 0.41| -0.16| 0    | 0    |
| ARHT     | 0.40            | 0.01       | 0.41     | 0.86| 0.44 | -0.04| 0.25| -0.25| 0.18 | 0    |
References

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.