Testing the independence of two random vectors where only one dimension is large

Weiming Li[1]

School of Science, Beijing University of Posts and Telecommunications, Beijing, China

Jiaqi Chen[2]

Department of Mathematics, Harbin Institute of Technology, Harbin, China

Jianfeng Yao

Department of Statistics and Actuarial Sciences, The University of Hong Kong, Hongkong, China

Abstract

For testing the independence of two vectors with respective dimensions p_1 and p_2, the existing literature in high-dimensional statistics all assume that both dimensions p_1 and p_2 grow to infinity with the sample size. However, as evidenced in RNA-sequencing data analysis, it happens frequently that one of the dimension is quite small and the other quite large compared to the sample size. In this paper, we address this new asymptotic framework for the independence test. A new test procedure is introduced and its asymptotic normality is established when the vectors are normally distributed. A Monte-Carlo study demonstrates the consistency of the procedure and ex-

Email addresses: liwm@bupt.edu.cn (Weiming Li), chenjq1016@gmail.com (Jiaqi Chen), jeffyao@hku.hk (Jianfeng Yao)

[1]Weiming Li's research is supported by National Natural Science Foundation of China, No. 11401037, and Fundamental Research Funds for the Central Universities, No. 2014RC0905.

[2]Jiaqi Chen's research is supported by Program for Innovation Research of Science in Harbin Institute of Technology, No. B201401.

hibits its superiority over some existing high-dimensional procedures. It is also shown that the procedure is robust against the normality assumption on the population vectors. Applied to a set of RNA-sequencing data, we obtain very convincing results on pairwise independence/dependence of gene isoform expressions as attested by prior knowledge established in that field.
Keywords: Covariance matrix, Gene network, High-dimensional testing, Independence test

1. Introduction

Modern scientific researches increasingly encounter high dimensional data and then evoke corresponding statistical analyses. In genomics, next-generation sequencing techniques such as RNA-Sequencing (Feng et al., 2013) are designed to quantify gene expression, where typically a group of gene isoforms are analyzed and their expression data at exon levels are recorded into multidimensional vectors. The dimensions of these vectors vary in a wide range where the smallest dimension can be one or two and the largest one can be comparable to the sample size. A fundamental issue in such analyses is determining whether there is any interaction between two given gene isoforms. More formally, this problem involves testing the independence of two possibly correlated vectors in a situation where one dimension is small but the other is large compared to the sample size.

Generally, let $\mathbf{X} = (X_1, \ldots, X_{p_1})$, $\mathbf{Y} = (Y_1, \ldots, Y_{p_2})$ and $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$ be the joint vector of dimension $p := p_1 + p_2$. The covariance matrix of $\mathbf{Z}$ is partitioned as

$$\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}$$

so that $\Sigma_{xx} = Var(\mathbf{X})$, $\Sigma_{yy} = Var(\mathbf{Y})$ and $\Sigma_{xy} = Cov(\mathbf{X}, \mathbf{Y})$. Let $\mathbf{z}_1, \ldots, \mathbf{z}_N$ be a sample of size N drawn from the population $\mathbf{Z}$. The sample covariance matrix is

$$S_n = \frac{1}{n} \sum_{k=1}^{N} (\mathbf{z}_k - \bar{\mathbf{z}})(\mathbf{z}_k - \bar{\mathbf{z}})'$$

where $\bar{\mathbf{z}} = \frac{1}{N} \sum_{k=1}^{N} \mathbf{z}_k$ and $n = N - 1$ represents the degree of freedom. Accordingly, S_n can be partitioned as

$$S_n = \begin{pmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{pmatrix}.$$

Assume that the joint vector $\mathbf{Z}$ has a p-dimensional normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ, the independence hypotheses of $\mathbf{X}$ and $\mathbf{Y}$ can be represented as

$$H_0 : \Sigma_{xy} = 0 \quad v.s. \quad H_1 : \Sigma_{xy} \neq 0. \tag{1}$$

To test these hypotheses, the following three statistics are commonly used (Anderson, 2003), which are the likelihood ratio test (LRT) and two trace criteria:

$$\Lambda = \frac{\sup_{H_0} L(\boldsymbol{\mu}, \Sigma)}{\sup L(\boldsymbol{\mu}, \Sigma)} = \frac{|S_n|^{N/2}}{|S_{xx}|^{N/2} |S_{yy}|^{N/2}} = |\mathbf{I}_{p_1} - S_{xy} S_{yy}^{-1} S_{yx} S_{xx}^{-1}|^{\frac{N}{2}},$$
$$C_1 = \mathrm{tr}(S_{xy} S_{yy}^{-1} S_{yx} S_{xx}^{-1}) \quad \text{and} \quad C_2 = \mathrm{tr}(S_{xy} S_{yx}) - \frac{1}{n} \mathrm{tr}(S_{xx}) \mathrm{tr}(S_{yy}) \tag{2}$$

The LRT statistic is the well-known Wilks's Λ (Wilks, 1935). Both statistics C_1 and C_2 are based on the idea that under the independence hypothesis, $\Sigma_{xy} = \Sigma'_{yx} = 0$ so that S_{xy} as well as S_{yx} should be small. A noticeable difference here is that the statistics Λ and C_1 rely on the inverse matrices S_{xx}^{-1} and S_{yy}^{-1} so that essentially the conditions $p_i < n$ are required. Conversely,

the criterion C_2 can be applied when the dimensions p_i, $i = 1, 2$, are larger than the sample size N.

The test procedures for the classical situation where the dimensions p_i's are reasonably small compared with the the sample size are well studied (Anderson, 2003). It is however well understood today that these asymptotical approximations are no more valid when the dimensions are comparable to the sample size, see e.g. Ledoit & Wolf (2002), Bai et al. (2009), Chen & Qin (2010) and Wang & Yao (2013). New limiting distributions have to be found in the large-dimensional context.

Specifically for the independence test, the existing literature in the large-dimensional context includes

1. the large-dimensional limit of Λ proposed in Jiang et al. (2013) under the asymptotic scheme $\min(p_1, p_2, n) \to \infty$, $p_1 + p_2 < n$ and $p_i/n \to c_i > 0$;
2. the large-dimensional limit of C_1 proposed in Jiang et al. (2013) under the asymptotic scheme $\min(p_1, p_2, n) \to \infty$, $\max(p_1, p_2) < n$ and $p_i/n \to c_i > 0$; and
3. the large-dimensional limit of C_2 proposed in Srivastava & Reid (2012) under the asymptotic scheme $\min(p_1, p_2, n) \to \infty$, $p_i/p \to d_i > 0$ and $n = O(p^\delta)$ for some constant $\delta > 0$ as $n \to \infty$.

Most recently, the trace criterion C_1 is generalized by Yang and Pan (2015) based on regularized canonical correlation coefficients to accommodate situations where $\max\{p_1, p_2\}$ can be larger than n. These existing asymptotic schemes are quite similar in that they all require that both dimensions p_1 and p_2 grow to infinity with the sample size N.

Motivated by RNA-sequencing analysis, our objective in this paper is

to test the hypotheses in (1) with the criterion C_2 assuming p_1 fixed and $\min(p_2, n) \to \infty$. As far as we know, this scheme has not been addressed in the literature. It will be proved that the asymptotic distribution of the statistic exists under this asymptotic scenario and is consistent with the one in Srivastava & Reid (2012). Note that our proof is different from theirs and this new asymptotic scenario is not covered by their results.

The rest of this paper is organised as follows. In the next section, we present the new test procedure and examine its size and power through simulation experiments. Section 3 presents an analysis of a genomic data set and Section 4 presents some conclusions and remarks. The main theorem is proved in the last section.

2. Test for the independence in high dimensions

2.1. Test statistic and its asymptotic distribution

The null hypothesis in (1) is equivalent to $\mathrm{tr}(\Sigma_{xy}\Sigma_{yx}) = 0$. Thus we may construct an unbiased estimator of this trace and reject the null hypothesis when this statistic is too large. Let

$$\gamma_2 = \mathrm{tr}(\Sigma^2), \quad \gamma_{xx} = \mathrm{tr}(\Sigma_{xx}^2), \quad \gamma_{yy} = \mathrm{tr}(\Sigma_{yy}^2), \quad \gamma_{xy} = \mathrm{tr}(\Sigma_{xy}\Sigma_{yx}).$$

We have by definition $2\gamma_{xy} = \gamma_2 - \gamma_{xx} - \gamma_{yy}$. From Srivastava (2005), an unbiased estimator of γ_2 is given as $k_n[\mathrm{tr}(S_n^2) - \mathrm{tr}^2(S_n)/n]$ with $k_n = n^2/(n-1)(n+2)$. Therefore an unbiased estimator of γ_{xy} is constructed as

$$\begin{aligned}\hat{\gamma}_{xy} &= \frac{k_n}{2}\left\{\mathrm{tr}(S_n^2) - \mathrm{tr}(S_{xx}^2) - \mathrm{tr}(S_{yy}^2) - \frac{1}{n}\left[\mathrm{tr}^2(S_n) - \mathrm{tr}^2(S_{xx}) - \mathrm{tr}^2(S_{yy})\right]\right\}, \\ &= k_n\left[\mathrm{tr}(S_{xy}S_{yx}) - \frac{1}{n}\mathrm{tr}(S_{xx})\mathrm{tr}(S_{yy})\right].\end{aligned}$$

We thus get the trace criterion C_2 given in (2). Notice that the estimator $\hat{\gamma}_{xy}$ is a function of eigenvalues of the sample covariance matrices S_{xx}, S_{yy}, and S_n.

Theorem 1. *Suppose that the dimension p_1 is fixed, p_2 and n both tend to infinity, and*

$$0 < \lim_{p\to\infty} \frac{1}{p}\mathrm{tr}(\Sigma^k) < \infty, \ \ k = 1, 2, 4.$$

Then under the null hypothesis in (1),

$$T_n := \frac{n}{\sqrt{2k_n}} \frac{\hat{\gamma}_{xy}}{\sqrt{\hat{\gamma}_{xx}\hat{\gamma}_{yy}}} \xrightarrow{d} N(0, 1), \tag{3}$$

where $\hat{\gamma}_{xx} = k_n[\mathrm{tr}(S_{xx}^2) - \mathrm{tr}^2(S_{xx})/n]$ and $\hat{\gamma}_{yy} = k_n[\mathrm{tr}(S_{yy}^2) - \mathrm{tr}^2(S_{yy})/n]$ with $k_n = n^2/(n-1)(n+2)$.

This theorem establishes the asymptotic distribution of T_n for the scenario that p_1 is fixed and p_2 approaches infinity. Notice that the convergence in (3) coincides with the result in Srivastava & Reid (2012), which assumes both p_1 and p_2 tend to infinity. This means that for practical applications, the proposed test is robust against different asymptotic scenarios of dimensions. Such robustness is especially welcomed since in a precise application (such as the gene isoform data analyzed in this paper) the explicit values of the dimensions p_1 and p_2 are known and it is somehow difficult to decide what is the most convenient asymptotic scenario to use. Synthesizing the two scenarios, we immediately get a more general one, i.e. $p = p_1 + p_2 \to \infty$, for the convergence in (3).

Theorem 2. *Suppose that the dimensions $p = p_1 + p_2$ and n both tend to infinity, and*

(a) $0 < \lim_{p\to\infty} \frac{1}{p}\mathrm{tr}(\Sigma^k) < \infty, \ \ k = 1, 2, 4;$

(b) *for some* $\delta \geq 0$, $\lim_{p,n\to\infty} \frac{p^{1-\delta}}{n} = 0$ *and* $0 < \lim_{p\to\infty} \frac{1}{p^{\delta}} \mathrm{tr}(\Sigma_{xy}\Sigma_{yx})$.

Then the asymptotic power of the proposed test tends to 1 as $N \to \infty$.

Theorem 2 presents a set of sufficient conditions for the consistency of the test T_n. Condition (a) is the same as used in Theorem 1, which claims the existence of the fourth moments of the population spectral distribution. Condition (b) describes the range of the asymptotic frameworks that one can use. Condition (c) clarifies the requirement of the amount of $\mathrm{tr}(\Sigma_{xy}\Sigma_{yx})$ for the consistency in a specified framework. For instance, in the most commonly used content $p/n \to c \in (0, \infty)$, we need $\mathrm{tr}(\Sigma_{xy}\Sigma_{yx}) \to \infty$ to guarantee the consistency.

2.2. Monte-Carlo study

We numerically evaluate the finite-sample performance of the test T_n and report the empirical size and power under different dimension settings. For the purpose of comparison, we also consider two tests discussed in Jiang et al. (2013): one is the corrected LRT, referred as T_1, and the other is based on the trace criterion C_1, referred as T_2. Since the test T_1 is limited to $p_1 + p_2 < n$ and T_2 is limited to $\max\{p_1, p_2\} < n$, we only consider the former case when comparing the three tests. The nominal significance level is fixed at $\alpha = 0.05$, and the number of independent replications is 100, 000.

We first report the empirical sizes of the three tests. Samples are drawn from standard normal population, and thus Σ is an identity matrix. The dimensions are $p_1 = 2, 6, 10$, $p_2 = 10, 30, 100, 200, 500$, and $n = 50$. The results are collected in Table 1, where the first six columns compare the sizes of the three tests when $p_1 + p_2 < n$ and the last three columns illustrate the

size of the proposed T_n when $p_2 > n$. The results show that all the empirical sizes are close to the nominal significance level.

Table 1: Empirical sizes in percents for the three tests with the significant level $\alpha = 0.05$.

p_1	$p_2 = 10$			$p_2 = 30$			T_n&p_2		
	T_n	T_1	T_2	T_n	T_1	T_2	100	200	500
2	6.32	6.56	5.86	5.72	6.17	4.48	5.52	5.34	5.30
6	5.89	6.11	5.37	5.66	5.88	4.69	5.46	5.21	5.29
10	5.74	6.03	5.27	5.46	5.90	4.70	5.36	5.09	5.16

To examine the powers of the three tests, we employ a model studied in Jiang et al. (2013), where the populations $\mathbf{X}$ and $\mathbf{Y}$ are defined as

$$\mathbf{X} = \mathbf{U}_1 + \gamma \mathbf{U}_2^{p_1}, \quad \mathbf{Y} = \mathbf{U}_2 + \gamma \mathbf{U}_2, \quad \mathbf{U}_i \sim N(0, \mathbf{I}_{p_i}), \quad i = 1, 2,$$

respectively, where $\mathbf{U}_1$ and $\mathbf{U}_2$ are independent, $\mathbf{U}_2^{p_1}$ is a subset of $\mathbf{U}_2$ consisting of its first p_1 variables, and the factor γ represents the degree of mixture. Therefore, the covariance matrices are respectively

$$\Sigma_{xx} = (1+\gamma^2) I_{p_1}, \quad \Sigma_{yy} = (1+\gamma)^2 I_{p_2}, \quad \Sigma_{xy} = \gamma(1+\gamma)(I_{p_1}, O_{p_1, p_2 - p_1}),$$

where $O_{m,n}$ represents an $m \times n$ zero matrix.

Figure 1 illustrates the powers of the three tests for this model. In the left panel, the parameters are $(p_1, p_2, n) = (4, 30, 50)$ and the factor γ increases from 0 to 0.9; while on the right, $(p_1, n, \gamma) = (4, 50, 0.5)$ and p_2 increases from 5 to 45. The curves in the figure show that the powers of the tests T_1 and T_2 are similar, and are dominated by the proposed test T_n in all the settings. Particularly, the curves in the right panel show that all the powers of the tests decrease as p_2 increases, which reflects the fact that in

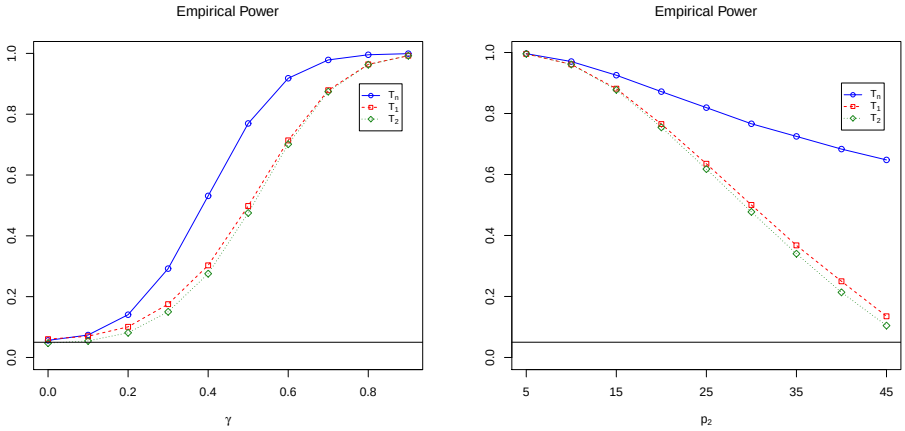

Figure 1: Empirical powers of the three tests. The parameter settings are $(p_1, p_2, n) = (4, 30, 50)$, $0 \leq \gamma \leq 0.9$ in the left panel, and $(p_1, n, \gamma) = (4, 50, 0.5)$, $5 \leq p_2 \leq 45$ in the right panel.

this process the increasing number of zero entries of Σ_{xy} makes it closer to the zero matrix of the null hypothesis. However, the power of T_n declines much slower than T_1 and T_2, which demonstrates a greater robustness of T_n against the inflating p_2.

Next we examine the robustness of the three test procedures when the assumed normal distribution of the vectors is contaminated by gamma-distributed errors. The studied model is the same as the previous except that the vector $\mathbf{U}_i$'s are replaced by

$$\mathbf{U}_i + \theta \mathbf{V}_i, \quad \mathbf{V}_i = (v_{i1}, \ldots, v_{ip_i})', \quad i = 1, 2,$$

where $\{v_{ij}\}$, independent of $\{\mathbf{U}_i\}$, are *i.i.d.* standardized random variables derived from $Gamma(a, b)$ distributed variables and the parameter θ represents the level of contamination. The new parameters are set to be $a = b = 3$ (positive skew, heavy-tailed) and $\theta = 1/2, 2$ in this experiment. Thus the

Table 2: Empirical sizes in percents for the three tests with the significant level $\alpha = 0.05$.

	p_1	$p_2 = 10$			$p_2 = 30$			T_n&p_2		
		T_n	T_1	T_2	T_n	T_1	T_2	100	200	500
$\theta = \frac{1}{2}$	2	6.38	6.56	5.90	5.86	6.19	4.45	5.55	5.32	5.22
	6	5.91	6.17	5.44	5.47	5.84	4.60	5.34	5.24	5.16
	10	5.71	5.94	5.18	5.55	5.80	4.80	5.33	5.12	5.22
$\theta = 2$	2	6.38	6.52	5.80	6.00	6.38	4.62	5.59	5.59	5.33
	6	6.02	6.15	5.49	5.65	5.79	4.67	5.42	5.42	5.22
	10	5.85	6.03	5.27	5.68	5.82	4.84	5.35	5.33	5.06

covariance matrices become

$$\Sigma_{xx} = (1+\gamma^2)(1+\theta^2)I_{p_1}, \quad \Sigma_{yy} = (1+\gamma)^2(1+\theta^2)I_{p_2},$$
$$\Sigma_{xy} = \gamma(1+\gamma)(1+\theta^2)(I_{p_1}, O_{p_1,p_2-p_1}).$$

Results about the empirical sizes and powers of the tests are collected in Table 2 and Figure 2, respectively. It shows that all the sizes are close to the nominal one and the power curves are quite similar to those in Figure 1, which demonstrate that the additional gamma-distributed errors have little impact on the three tests. It is however worth noticing that the theoretic proof of Theorem 1 in this paper as well as the proofs for asymptotic normality of the test criteria T_1 and T_2 established in Jiang et al. (2013) all heavily rely on the assumed normality of the vectors, and to our best knowledge, it seems unclear how these proofs can be extended to cover non-normal data as the ones tested in the Monte-Carlo experiments reported here.

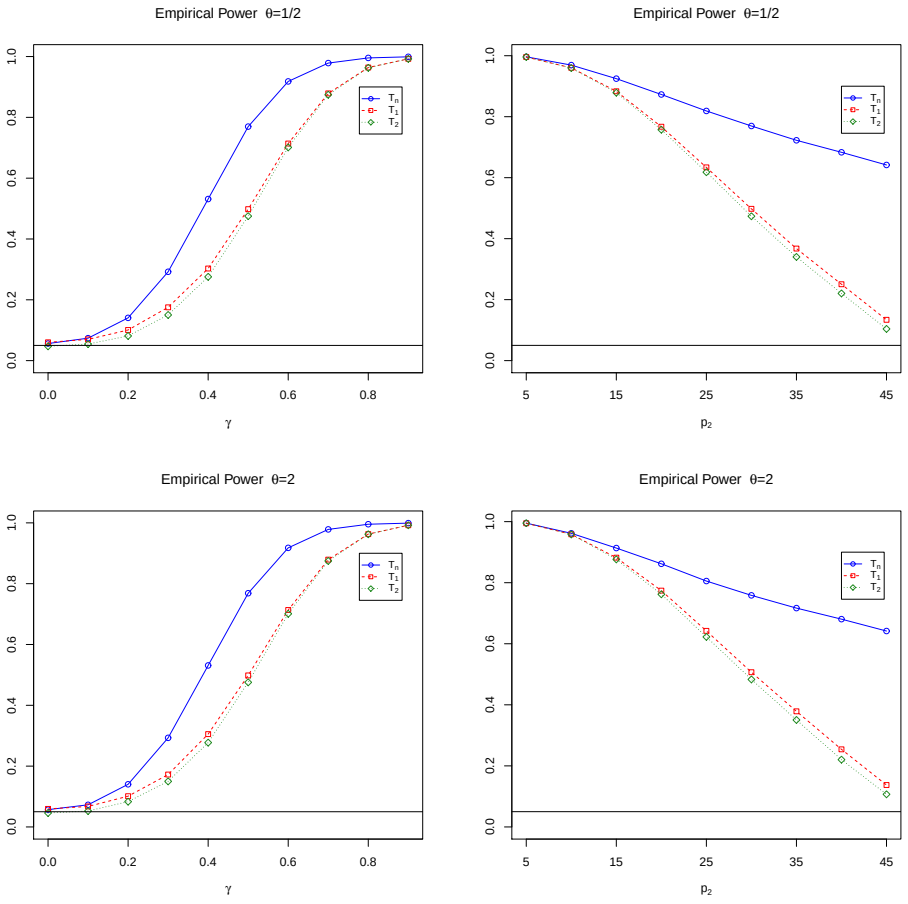

Figure 2: Empirical powers of the three tests for the non-normal distribution with $\theta = 1/2, 2$. The parameter settings are $(p_1, p_2, n) = (4, 30, 50)$, $0 \leq \gamma \leq 0.9$ in the left panel, and $(p_1, n, \gamma) = (4, 50, 0.5)$, $5 \leq p_2 \leq 45$ in the right panel.

3. Real data analysis

Genomes play a central role in the control of cellular processes (Barabasi & Oltvai, 2004). The dynamic interplay between various genes can be mapped as gene co-expression networks, which is an important and widely used method to understand the cause and prognosis of various diseases. To recover pairwise dependencies in a gene co-expression network, each co-expression edge has to be inferred by accepting or rejecting the independence hypothesis from the sample covariance matrix of respective isoform expressions.

We analyze a data set of liver cancer, which is downloaded from TCGA data portal: https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm, and filtered by data types RNASeqV2 and Level 3. The data set consists of 38 genes with their dimensions ranging from 1 to 31 (see Table 3) and their sample size is $N = 50$. Obviously the dimensions are not on the same order of magnitude as their sample size. For these genes, the relationship of dependency are totally known based on established knowledge from historical experiments: 29 pairs of them are dependent and the remaining 674 pairs are independent.

We test the pairwise gene dependencies using T_n and compare the results with those from two other methods: one is from Hong et al. (2013), which is a variant of traditional canonical correlation analysis (CCA); the other is the large-dimensional trace criterion T_2, which is recently applied in Yalamanchili et al. (2014) and is demonstrated better than CCA. The corrected LRT T_1 is excluded from this comparison since its dimensional requirement is not met for the data set. The significance level is set to be $\alpha = 0.05$. To evaluate the accuracy of the test results, we employ the so called F-score

Table 3: Lung cancer data: 38 genes with different dimensions

Name	NM000222	NM000321	NM000636	NM000791	NM001126116	NM001140
Dimension	20	27	4	6	7	13
Name	NM001145102	NM001204191	NM001237	NM001429	NM001759	NM001760
Dimension	9	7	8	31	5	4
Name	NM001786	NM001880	NM001950	NM002198	NM002228	NM002421
Dimension	4	13	10	9	1	10
Name	NM002467	NM002505	NM002539	NM002985	NM003109	NM003153
Dimension	3	10	11	3	6	22
Name	NM003221	NM003998	NM004379	NM004417	NM005194	NM005238
Dimension	7	23	8	2	1	8
Name	NM005239	NM005252	NM005438	NM007122	NM022457	NM033285
Dimension	10	2	4	11	20	4
Name	NM053056	NM198253				
Dimension	5	15				

(Powers, 2007) which actually measures the trade-off between precision P and recall R:

$$F = 2 \times \frac{P \times R}{P + R}, \tag{4}$$

where

$$P = \frac{true\ positives}{true\ positives + false\ positives}, \quad R = \frac{true\ positives}{true\ positives + false\ negatives}.$$

With the prior information of dependency, the *true positives* stands for the number of correctly identified correlated pairs of genes, the *false positive* is the number of misidentified correlated pairs of genes, and the *false negatives* is the number of misidentified uncorrelated pairs of genes.

The F-scores reported in Table 4 show that T_n outperforms T_2 significantly. CCA fails to detect the relationship between gene NM002228 and

other genes due to the dimension of this gene is 1. The same phenomenon happens to gene NM005195. Therefore, we cannot get F-score for CCA.

Table 4: F-scores for the data set including 38 genes.

Method	T_n	T_2	CCA
F-score	0.64	0.40	NA

Next, we remove the 1-dimensional genes from the data set in order to incorporate CCA for comparison. The remaining 36 genes include 25 dependent pairs and 605 independent pairs. The F-scores collected in Table 5 demonstrate that T_n again outperforms the others. Notice that such results on pairwise dependence of gene isoform expressions are further used to construct gene co-expression networks, see Yalamanchili et al. (2014).

Table 5: F-scores for the data set including 36 genes.

Method	T_n	T_2	CCA
F-score	0.6465	0.4238	0.4187

4. Concluding remarks

This paper investigates the independence test of two vectors in a high-dimensional situation where one of the dimensions p_1 is quite small while the other dimension p_2 is large compared to the sample size. The asymptotic scheme is novel and practically useful. A new procedure is introduced and the test statistic under the null is proved to be asymptotically normal distributed assuming that $p_1 + p_2 \to \infty$ and the vectors are normal distributed. The power of the proposed test is studied through Monte-Carlo simulations and a real data analysis, which demonstrates the superiority of the new test

over the existing ones. Another interesting feature found in the Monte-Carlo study is that the proposed procedure is robust against deviations from the normality assumption on the vectors although a theoretic proof of this fact is still missing.

5. Proofs

5.1. Lemma

Lemma 1. *Let* $\mathbf{u}$, $\mathbf{v}$, *and* $\mathbf{w}$ *be independent vectors of* n*-dimensional standard normal distribution* $N(0, I_n)$, *and define*

$$\psi(\mathbf{x}, \mathbf{y}) = \frac{1}{n}(\mathbf{x}'\mathbf{y})^2 - \frac{1}{n^2}(\mathbf{x}'\mathbf{x})(\mathbf{y}'\mathbf{y}), \tag{5}$$

then

$$\begin{aligned}
&\mathrm{E}[\psi(\mathbf{u}, \mathbf{v})|\mathbf{u}] = 0, \quad \mathrm{E}[\psi(\mathbf{u}, \mathbf{v})\psi(\mathbf{w}, \mathbf{v})|\mathbf{u}, \mathbf{w}] = \frac{2}{n}\psi(\mathbf{u}, \mathbf{w}), \\
&\mathrm{E}[\psi(\mathbf{v}, \mathbf{v})] = (n-1)(n+2)/n, \quad \mathrm{E}[\psi^2(\mathbf{u}, \mathbf{v})] = 2(n-1)(n+2)/n^2, \\
&\mathrm{E}[\psi^2(\mathbf{v}, \mathbf{v})] = O(n^2), \quad \mathrm{Var}[\psi^2(\mathbf{v}, \mathbf{v})] = O(n), \quad \mathrm{E}[\psi^4(\mathbf{u}, \mathbf{v})] = O(1),
\end{aligned}$$

as $n \to \infty$.

Proof. The distribution of $\mathbf{v}'\mathbf{v}$ is $\chi^2(n)$ and the conditional distribution of $\mathbf{u}'\mathbf{v}|\mathbf{u}$ is $N(0, \mathbf{u}'\mathbf{u})$, thus $\mathrm{E}[\psi(\mathbf{u}, \mathbf{v})|\mathbf{u}] = 0$. Write

$$\begin{aligned}
\psi(\mathbf{u}, \mathbf{v})\psi(\mathbf{w}, \mathbf{v}) &= \frac{1}{n^2}(\mathbf{u}'\mathbf{v})^2(\mathbf{w}'\mathbf{v})^2 - \frac{1}{n^3}(\mathbf{u}'\mathbf{v})^2(\mathbf{w}'\mathbf{w})(\mathbf{v}'\mathbf{v}) \\
&\quad -\frac{1}{n^3}(\mathbf{w}'\mathbf{v})^2(\mathbf{u}'\mathbf{u})(\mathbf{v}'\mathbf{v}) + \frac{1}{n^4}(\mathbf{u}'\mathbf{u})(\mathbf{w}'\mathbf{w})(\mathbf{v}'\mathbf{v})^2 \\
&:= \frac{1}{n^2}S_1 - \frac{1}{n^3}S_2 - \frac{1}{n^3}S_3 + \frac{1}{n^4}S_4.
\end{aligned}$$

Then $\mathrm{E}(S_4|\mathbf{u},\mathbf{w}) = n(n+2)(\mathbf{u}'\mathbf{u})(\mathbf{w}'\mathbf{w})$, and

$$\begin{aligned}
\mathrm{E}(S_1|\mathbf{u},\mathbf{w}) &= \sum_{i,j,k,l} u_i u_j w_k w_l \mathrm{E}(v_i v_j v_k v_l) \\
&= \sum_{i=j,k=l} u_i u_j w_k w_l + \sum_{i=k,j=l} u_i u_j w_k w_l + \sum_{i=l,j=k} u_i u_j w_k w_l \\
&= (\mathbf{u}'\mathbf{u})(\mathbf{w}'\mathbf{w}) + 2(\mathbf{u}'\mathbf{w})^2, \\
\mathrm{E}(S_2|\mathbf{u},\mathbf{w}) &= (\mathbf{w}'\mathbf{w}) \sum_{i,k} u_i^2 \cdot \mathrm{E}(v_i^2 v_k^2) = (n+2)(\mathbf{u}'\mathbf{u})(\mathbf{w}'\mathbf{w}),
\end{aligned}$$

and thus $\mathrm{E}(S_3|\mathbf{u},\mathbf{w}) = \mathrm{E}(S_2|\mathbf{u},\mathbf{w})$, where (x_i) denote the elements of $\mathbf{x}$. Collecting these results, we get $\mathrm{E}[\psi(\mathbf{u},\mathbf{v})\psi(\mathbf{w},\mathbf{v})|\mathbf{u},\mathbf{w}] = (2/n)\psi(\mathbf{u},\mathbf{w})$.

Notice that $\psi(\mathbf{v},\mathbf{v}) = (n-1)(\mathbf{v}'\mathbf{v})^2/n^2$, and $\mathrm{E}(\mathbf{v}'\mathbf{v})^k = n(n+2)\cdots(n+2k-2)$, $k \in \mathrm{N}^+$. We have then,

$$\begin{aligned}
\mathrm{E}[\psi(\mathbf{v},\mathbf{v})] &= (n-1)(n+2)/n, \\
\mathrm{E}[\psi^2(\mathbf{u},\mathbf{v})] &= (2/n)\mathrm{E}[\psi(\mathbf{u},\mathbf{u})] = 2(n-1)(n+2)/n^2, \\
\mathrm{E}[\psi^2(\mathbf{v},\mathbf{v})] &= \mathrm{E}(\mathbf{v}'\mathbf{v})^4 (n-1)^2/n^4 = O(n^2), \\
\mathrm{Var}[\psi^2(\mathbf{v},\mathbf{v})] &= \left[\mathrm{E}(\mathbf{v}'\mathbf{v})^4 - \mathrm{E}^2(\mathbf{v}'\mathbf{v})^2\right](n-1)^2/n^4 = O(n).
\end{aligned}$$

Finally, from Minkowski inequality,

$$\begin{aligned}
\mathrm{E}[\psi^4(\mathbf{u},\mathbf{v})] &= \frac{1}{n^4}\mathrm{E}[(\mathbf{u}'\mathbf{v})^2 - (\mathbf{u}'\mathbf{u})(\mathbf{v}'\mathbf{v})/n]^4 \\
&\leq \frac{1}{n^4}\left\{\left[\mathrm{E}(\mathbf{u}'\mathbf{v})^8\right]^{\frac{1}{4}} + \left[\mathrm{E}(\mathbf{u}'\mathbf{u})^4\mathrm{E}(\mathbf{v}'\mathbf{v})^4\right]^{\frac{1}{4}}/n\right\}^4 \\
&= \frac{1}{n^4}\left\{\left[\mathrm{E}(\mathbf{v}'\mathbf{v})^4\right]^{\frac{1}{4}} + \left[\mathrm{E}(\mathbf{v}'\mathbf{v})^4\right]^{\frac{1}{2}}/n\right\}^4,
\end{aligned}$$

which is $O(1)$ as $n \to \infty$. □

5.2. Proof of Theorem 1

The sample covariance S_n has the Wishart distribution $W_n(\Sigma)$ with n degrees of freedom. It can be expressed as $\sum_{i=1}^n \tilde{\mathbf{z}}_k \tilde{\mathbf{z}}_k'/n$ where $(\tilde{\mathbf{z}}_i)$ are i.i.d.

$N(0,\Sigma)$. Write $\tilde{\mathbf{z}}_i = (\tilde{\mathbf{x}}_i', \tilde{\mathbf{y}}_i')' = (\tilde{x}_{i1}, \ldots, \tilde{x}_{ip_1}, \tilde{y}_{i1}, \ldots, \tilde{y}_{ip_2})'$, $i = 1, \ldots n$, and denote $\mathbf{X} = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n)$ and $\mathbf{Y} = (\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_n)$. Note that the matrices $\mathbf{X}$ and $\mathbf{Y}$ contain normal vectors which are independent under H_0. The matrices $\mathbf{X}'\mathbf{X}$ and $\mathbf{Y}'\mathbf{Y}$ can be standardized as

$$\mathbf{X}'\mathbf{X} = \sum_{i=1}^{p_1} \alpha_i \mathbf{u}_i \mathbf{u}_i', \quad \mathbf{Y}'\mathbf{Y} = \sum_{j=1}^{p_2} \beta_j \mathbf{v}_j \mathbf{v}_j',$$

where (α_i) and (β_j) are the eigenvalues of Σ_{xx} and Σ_{yy}, respectively, and $(\mathbf{u}_i), (\mathbf{v}_j)$ are i.i.d. $N(0, I_n)$. Therefore, we have

$$\begin{aligned}
\frac{n}{k_n}\hat{\gamma}_{xy} &= n\mathrm{tr}(S_{xy}S_{yx}) - \mathrm{tr}(S_{xx})\mathrm{tr}(S_{yy}) \\
&= \frac{1}{n}\mathrm{tr}(\mathbf{X}'\mathbf{X}\mathbf{Y}'\mathbf{Y}) - \frac{1}{n^2}\mathrm{tr}(\mathbf{X}'\mathbf{X})\mathrm{tr}(\mathbf{Y}'\mathbf{Y}) \\
&= \sum_{i=1}^{p_1}\sum_{j=1}^{p_2} \alpha_i\beta_j \left[\frac{1}{n}(\mathbf{u}_i'\mathbf{v}_j)^2 - \frac{1}{n^2}(\mathbf{u}_i'\mathbf{u}_i)(\mathbf{v}_j'\mathbf{v}_j)\right] \\
&:= \sum_{i=1}^{p_1}\sum_{j=1}^{p_2} a_{ij}\psi(\mathbf{u}_i, \mathbf{v}_j),
\end{aligned}$$

where $a_{ij} = \alpha_i\beta_j$ and ψ is defined in (5) with the dimension n, $i = 1, \ldots, p_1$, $j = 1, \ldots, p_2$.

We use the martingale CLT to establish the limiting distribution of T_n. Without loss of generality suppose that $p_1 \leq p_2$, and define $\phi_j^{(n)} = (1/\sqrt{p_1 p_2}) \sum_{i=1}^{p_1} a_{ij}\psi(\mathbf{u}_i, \mathbf{v}_j)$, $j = 1, \ldots, p_2$. Let $\mathcal{F}_j^{(n)}$ be the σ-algebra generated by the random variables $\{\mathbf{u}_1, \ldots, \mathbf{u}_{p_1}, \mathbf{v}_1, \ldots, \mathbf{v}_j\}$, then $\{\emptyset, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1^{(n)} \subset \cdots \subset \mathcal{F}_{p_2}^{(n)} \subset \mathcal{F}$ with $(\Omega, \mathcal{F}, P)$ the probability space. From Lemma

1 and the law of iterated expectations,

$$\begin{aligned}
\mathrm{E}\left[\phi_j^{(n)}\middle|\mathcal{F}_{j-1}^{(n)}\right] &= \frac{1}{\sqrt{p_1 p_2}}\sum_{i=1}^{p_1} a_{ij}\mathrm{E}(\psi(\mathbf{u}_i,\mathbf{v}_j)|\mathbf{u}_i) = 0,\\
\mathrm{E}\left[\phi_j^{(n)}\right]^2 &= \frac{1}{p_1 p_2}\sum_{i=1}^{p_1}\sum_{k=1}^{p_1} a_{ij}a_{kj}\mathrm{E}[\psi(\mathbf{u}_i,\mathbf{v}_j)\psi(\mathbf{u}_k,\mathbf{v}_j)]\\
&= \frac{2(n-1)(n+2)}{n^2 p_1 p_2}\sum_{i=1}^{p_1} a_{ij}^2,
\end{aligned}$$

which is $O(1/p_2)$ as $(p,n)\to\infty$. Thus $\{\psi_j^{(n)},\mathcal{F}_j^{(n)}\}$ forms a sequence of integrable martingale differences. On the other hand,

$$\begin{aligned}
\sum_{j=1}^{p_2}\mathrm{E}\left[\left(\phi_j^{(n)}\right)^2\middle|\mathcal{F}_{j-1}^{(n)}\right] &= \frac{1}{p_1 p_2}\sum_{j=1}^{p_2}\sum_{i=1}^{p_1}\sum_{k=1}^{p_1} a_{ij}a_{kj}\mathrm{E}\left(\psi(\mathbf{u}_i,\mathbf{v}_j)\psi(\mathbf{u}_k,\mathbf{v}_j)|\mathbf{u}_i,\mathbf{u}_k\right)\\
&= \frac{2}{np_1 p_2}\sum_{i=1}^{p_1}\sum_{k=1}^{p_1} b_{ij}\psi(\mathbf{u}_i,\mathbf{u}_k)\\
&= \frac{2}{np_1 p_2}\sum_{i=1}^{p_1} b_{ii}\psi(\mathbf{u}_i,\mathbf{u}_i) + \frac{2}{np_1 p_2}\sum_{i\neq k} b_{ik}\psi(\mathbf{u}_i,\mathbf{u}_k)\\
&:= A_{1n} + A_{2n},
\end{aligned}$$

where $b_{ik} = \sum_{j=1}^{p_2} a_{ij}a_{kj}$, $i,k = 1,\ldots,p_1$. Considering the variances of A_{1n} and A_{2n}, $\mathrm{Var}(A_{1n}) = O(1/n)$ and

$$\begin{aligned}
\mathrm{Var}(A_{2n}) &= \frac{4}{n^2 p_1^2 p_2^2}\sum_{i\neq k}\sum_{l\neq s} b_{ik}b_{ls}\mathrm{E}[\psi(\mathbf{u}_i,\mathbf{u}_k)\psi(\mathbf{u}_l,\mathbf{u}_s)]\\
&= \frac{8}{n^2 p_1^2 p_2^2}\sum_{i\neq k} b_{ik}^2\mathrm{E}[\psi^2(\mathbf{u}_i,\mathbf{u}_k)],
\end{aligned}$$

which is $O(1/n^2)$. Therefore, from the Chebyshev inequality,

$$\sum_{j=1}^{p_2}\mathrm{E}\left[\left(\phi_j^{(n)}\right)^2\middle|\mathcal{F}_{j-1}^{(n)}\right] - \sum_{j=1}^{p_2}\mathrm{E}\left(\phi_j^{(n)}\right)^2 \xrightarrow{p} 0, \quad \text{as} \quad (p,n)\to\infty,$$

where the second expectation has expression $s_n^2 := 2(1-1/n)(1+2/n)\sum_{i=1}^{p_1}\sum_{j=1}^{p_2} a_{ij}^2/(p_1 p_2)$.

Next we verify Lyapunov condition by showing that $B_n = \sum_{j=1}^{p_2} \mathrm{E}(\phi_j^{(n)})^4 \to 0$. From Lemma 1 and the law of iterated expectations,

$$\begin{aligned} B_n &= \frac{1}{p_1^2 p_2^2} \sum_{j=1}^{p_2} \sum_{i=1}^{p_1} \sum_{l=1}^{p_1} \sum_{s=1}^{p_1} \sum_{t=1}^{p_1} a_{ij} a_{lj} a_{sj} a_{tj} \mathrm{E}[\psi(\mathbf{u}_i, \mathbf{v}_j)\psi(\mathbf{u}_l, \mathbf{v}_j)\psi(\mathbf{u}_s, \mathbf{v}_j)\psi(\mathbf{u}_t, \mathbf{v}_j)] \\ &= \frac{1}{p_1^2 p_2^2} \sum_{j=1}^{p_2} \sum_{i=1}^{p_1} a_{ij}^4 \mathrm{E}[\psi^4(\mathbf{u}_i, \mathbf{v}_j)] + \frac{3}{p_1^2 p_2^2} \sum_{j=1}^{p_2} \sum_{i \neq s} a_{ij}^2 a_{sj}^2 \mathrm{E}[\psi^2(\mathbf{u}_i, \mathbf{v}_j)\psi^2(\mathbf{u}_s, \mathbf{v}_j)] \\ &= \frac{1}{p_1^2 p_2^2} \sum_{j=1}^{p_2} \sum_{i=1}^{p_1} a_{ij}^4 \mathrm{E}[\psi^4(\mathbf{u}_i, \mathbf{v}_j)] + \frac{12}{n^2 p_1^2 p_2^2} \sum_{j=1}^{p_2} \sum_{i \neq s} a_{ij}^2 a_{sj}^2 \mathrm{E}[\psi^2(\mathbf{v}_j, \mathbf{v}_j)], \end{aligned}$$

which is $O(1/p_2)$ as $(p, n) \to \infty$.

Notice that $\hat{\gamma}_{xx}$ and $\hat{\gamma}_{yy}$ are unbiased and consistent estimators of γ_{xx} and γ_{yy}, respectively. The statistic $\hat{s}_n^2 := 2(1 - 1/n)(1 + 2/n)\hat{\gamma}_{xx}\hat{\gamma}_{yy}/(p_1 p_2)$ is also an unbiased and consistent estimator of s_n^2 under the null hypothesis, therefore

$$\frac{n}{\sqrt{2k_n}} \frac{\hat{\gamma}_{xy}}{\sqrt{\hat{\gamma}_{xx}\hat{\gamma}_{yy}}} = \frac{1}{\hat{s}_n} \sum_{j=1}^{p_2} \phi_j^{(n)} \xrightarrow{d} N(0, 1), \quad \text{as} \quad (p, n) \to \infty.$$

5.3. Proof of Theorem 2

Denote $\hat{\gamma}_2 = k_n[\mathrm{tr}(S_n^2) - \frac{1}{n}\mathrm{tr}^2(S_n)]$. Under the assumptions of the theorem, from Lemma 6.4 in Srivastava (2005), we have $E(\hat{\gamma}_2) = \mathrm{tr}(\Sigma^2)$ and

$$Var(\hat{\gamma}_2) = \frac{8(n+3)}{n(n+2)}\mathrm{tr}(\Sigma^4) + \frac{4}{(n+2)(n-1)}\left(\mathrm{tr}^2(\Sigma^2) - \mathrm{tr}(\Sigma^4)\right).$$

From Chebyshev inequality, $(\hat{\gamma}_2 - \mathrm{tr}(\Sigma^2))/p$ converges in probability to 0, as n, p tend to infinity. Similarly, we can get

$$\frac{1}{p_1}(\hat{\gamma}_{xx} - \gamma_{xx}) \xrightarrow{p} 0, \quad \frac{1}{p_2}(\hat{\gamma}_{yy} - \gamma_{yy}) \xrightarrow{p} 0. \tag{6}$$

For the statistic $\hat{\gamma}_{xy}$, it's easy to see that $\hat{\gamma}_{xy} = (\hat{\gamma}_2 - \hat{\gamma}_{xx} - \hat{\gamma}_{yy})/2$, and thus we get $E(\hat{\gamma}_{xy}) = \gamma_{xy}$ and

$$\begin{aligned} Var(\hat{\gamma}_{xy}/p^{\delta}) &= \frac{1}{4p^{\delta 2}} Var(\hat{\gamma}_2 - \hat{\gamma}_{xx} - \hat{\gamma}_{yy}) \\ &\leq \frac{3}{4p^{\delta 2}} \left(Var(\hat{\gamma}_2) + Var(\hat{\gamma}_{xx}) + Var(\hat{\gamma}_{yy})\right) \to 0. \end{aligned}$$

Therefore, $(\hat{\gamma}_{xy} - \gamma_{xy})/p^{\delta} \xrightarrow{p} 0$, as n, p tend to infinity. From this, the results in (6), and again the assumptions of the theorem, we conclude that

$$T_n = \frac{n}{p^{1-\delta}} \cdot \frac{p}{\sqrt{2k_n p_1 p_2}} \cdot \frac{\hat{\gamma}_{xy}/p^{\delta}}{\sqrt{\hat{\gamma}_{xx}\hat{\gamma}_{yy}/p_1 p_2}} \xrightarrow{p} \infty, \quad n, p \to \infty,$$

which implies the result of the theorem.

References

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis*, third ed. Wiley & Sons, Hoboken, NJ.

Bai, Z. D., Jiang, D. D., Yao, J. F., and Zheng, S. R. (2009). Corrections to LRT on large-dimensional covariance matrix by RMT. *Ann. Statist.*, **37**, 3822–3840.

Barabasi, A. L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature reviews genetics*, **5**, 101–113.

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

Feng, H., Qin, Z., and Zhang, X. (2013). Opportunities and methods for studying alternative splicing in cancer with RNA-Seq. *Cancer lett.*, **340**, 179–191.

Hong, S., Chen, X., Jin, L., and Xiong, M. (2013). Canonical correlation analysis for RNA-seq co-expression networks. *Nucleic Acids Res.*, **41**, e95.

Jiang, D. D., Bai, Z. D., and Zheng, S. R. (2013). Testing the independence of sets of large-dimensional variables. *Sci. China Math.*, **56**, 135–147.

Ledoit, O. and Wolf, M. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Ann. Statist.*, **30**, 1081–1102.

Powers, D. W. M. (2007). Evaluation: from precision, recall and F-factor to ROC, informedness, markedness & correlation. *J. Mach. Learn. Technol.*, **2**, 37–63.

Wilks, S. S. (1935). On the independence of k sets of normally distributed statistical variables. *Econometrica*, **3**, 309–326.

Srivastava, M. S. (2005). Some tests concerning the covariance matrix in high dimensional data. *J. Japan Statist. Soc.*, **35**, 251–272.

Srivastava, M. S. and Reid, N. (2012). Testing the structure of the covariance matrix with fewer observations than the dimension. *J. Multivariate Anal.*, **112**, 156–171.

Wang, Q. W. and Yao, J. F. (2013). On the sphericity test with large-dimensional observations. *Electron. J. Stat.*, **7**, 2164–2192.

Yalamanchili, H. K., Li, Z. Y., Wang, P., Wong, M. P., Yao, J. F., and Wang, J. (2014). SpliceNet: recovering splicing isoform-specific differential gene networks from RNA-Seq data of normal and diseased samples. *Nucleic Acids Res.*, doi: 10.1093/nar/gku577.

Yang, Y. R. and Pan, G. M. (2015). Independence test for high dimensional data based on regularized canonical correlation coefficients. *Ann. Statist.*, **43**, 467–500.