On a test for assessing vector correlation for latent factor models in high-dimensional settings

Takahiro Nishiyama[a], Masashi Hyodo[b] and Shoichi Narita[c]

[a] Department of Business Administration, Senshu University
[b] Faculty of Economics, Kanagawa University
[c] Graduate School of Economics, Kanagawa University

1. Introduction

We let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be p-dimensional random sample with a population mean vector $\boldsymbol{\mu}$ and population covariance matrix $\boldsymbol{\Sigma}$. We further partition $\mathbf{x}_i$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ into 2 components:

$$\mathbf{x}_i = \begin{pmatrix} \mathbf{x}_{1i} \\ \mathbf{x}_{2i} \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \ \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

where $\mathbf{x}_{gi}$ and $\boldsymbol{\mu}_g$ are $p_g \times 1$ vectors, and $\boldsymbol{\Sigma}_{gh}$ is a $p_g \times p_h$ matrix, $g, h \in \{1, 2\}$. Note that $p = p_1 + p_2$. The test for assessing the vector correlation can be fomulated as

$$\mathcal{H} : \boldsymbol{\Sigma}_{12} = \mathbf{O} \ \text{ vs. } \ \mathcal{A} : \boldsymbol{\Sigma}_{12} \neq \mathbf{O}. \tag{1}$$

To construct test (1), we introduce the ρV coefficient introduced in [2]. The ρV coefficient of $\mathbf{x}_{1i}$ and $\mathbf{x}_{2i}$ is defined as

$$\rho V_{12} = \frac{\|\boldsymbol{\Sigma}_{12}\|_F^2}{\|\boldsymbol{\Sigma}_{11}\|_F \|\boldsymbol{\Sigma}_{22}\|_F},$$

where $\| \cdot \|_F$ denotes the Frobenius norm. The ρV-coefficient measures the correlation between two probability vectors. Particularly , if $p_1 = p_2 = 1$, it corresponds to the square of Pearson's correlation coefficient. Because $\boldsymbol{\Sigma}_{12} = \mathbf{O}$ and $\rho V_{12} = 0$ are equivalent, the estimator of ρV_{12} can be used to hypothesize testing (1). The RV coefficient introduced by [4] can be interpreted as a naive estimator of ρV-coefficient. However, [3] states that the RV coefficient takes high values when the sample size n is small, and when both p_1 and p_2 are large. Further, they corrected the RV coefficient so that it is consistent even in high-dimensional settings, and showed the asymptotic normality of the corrected RV under a high-dimensional framework with a multivariate normal population and the following covariance structure: (hereafter referred to as weak-spike structure).

$$\frac{\|\boldsymbol{\Sigma}_{gg}^2\|_F^2}{\|\boldsymbol{\Sigma}_{gg}\|_F^4} = o(1) \quad (p \to \infty). \tag{2}$$

This study provides ρV-based test for (1) without the normality assumption and weak-spike structure (2), while allowing the dimension p to be much larger than the sample size n.

2. Meain results

2.1. Data generation model and asymptotic framework

The data generation model is assumed to be a latent factor model expressed as

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{B}\mathbf{f} + \boldsymbol{\epsilon}. \tag{3}$$

Here, $\boldsymbol{\mu} \in \mathbb{R}^p$ is the population mean vector, $\mathbf{B}$ is the $p \times d$ non-random matrix $\mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_p)^\top$ that satisfies $\mathrm{rank}(\mathbf{B}) = d$, and elements $\mathbf{b}_1, \dots, \mathbf{b}_p$ are referred to as factor loadings. $\mathbf{f} \in \mathbb{R}^d$ and $\boldsymbol{\epsilon} \in \mathbb{R}^p$ are random vectors for common and specific factors, respectively. We assume that $\mathbf{f}$ and $\boldsymbol{\epsilon}$ are independent. We let $\mathbf{f} = (\mathrm{f}_1, \dots, \mathrm{f}_d)$ and $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_p)^\top$. Furthermore, we assume that f_i is iid with $\mathrm{E}(\mathrm{f}_i) = 0$, $\mathrm{E}(\mathrm{f}_i^2) = 1$, and $\mathrm{E}(\mathrm{f}_i^4) = \kappa + 3 < \infty$. and ϵ_j are iid with $\mathrm{E}(\epsilon_j) = 0$, $0 < \mathrm{E}(\epsilon_j^2) = \psi_j < \infty$, $\mathrm{E}(\epsilon_j^4) = \psi_j^2(\kappa + 3) < \infty$ for $i \in \{1, \dots, d\}$, and $j \in \{1, \dots, p\}$. Under these assumptions, $\mathrm{E}(\mathbf{f}) = \mathbf{0}$, $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$, $\mathrm{cov}(\mathbf{f}) = \mathbf{I}_d$ and $\mathrm{cov}(\boldsymbol{\epsilon}) = \boldsymbol{\Psi} = \mathrm{diag}(\psi_1, \dots, \psi_p)$.

We further partition $\mathbf{B}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\epsilon}$ into 2 components:

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix}, \ \boldsymbol{\Psi} = \begin{pmatrix} \boldsymbol{\Psi}_1 & \mathbf{O} \\ \mathbf{O} & \boldsymbol{\Psi}_2 \end{pmatrix}, \ \boldsymbol{\epsilon} = \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \end{pmatrix},$$

where $\mathbf{B}_g$ is $p_g \times d$ nonrandom matrix that satisfies $\mathrm{rank}(\mathbf{B}_g) = d_g > 0$, $\boldsymbol{\Psi}_g$ is $p_g \times p_g$ diagonal matrix, and $\boldsymbol{\epsilon}_g$ is p_g-dimensional random vector. These assumptions, along with Equation (3), imply that

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \ \boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}^\top + \boldsymbol{\Psi} = \begin{pmatrix} \mathbf{B}_1\mathbf{B}_1^\top + \boldsymbol{\Psi}_1 & \mathbf{B}_1\mathbf{B}_2^\top \\ \mathbf{B}_2\mathbf{B}_1^\top & \mathbf{B}_2\mathbf{B}_2^\top + \boldsymbol{\Psi}_2 \end{pmatrix}.$$

For the asymptotic evaluation, we impose the following regularity conditions:

(A1) $p_g = p_g(n)$ $(g \in \{1, 2\})$ is a function of n such that p_g tends to infinity along with $n \to \infty$, $n/p_g \to \theta_g \in (0, \infty)$, and positive integer d is fixed.

(A2) $\psi_{\max} = \max\{\psi_1, \psi_2, \dots, \psi_p\}$ is bounded.

(A3) There are two positive semidefinite matrices $\mathbf{B}_{11}^*$ and $\mathbf{B}_{22}^*$ such that $\mathrm{rank}(\mathbf{B}_{11}^*) = d_1 > 0$, $\mathrm{rank}(\mathbf{B}_{22}^*) = d_2 > 0$, and $\|(1/p_g)\mathbf{B}_g^\top\mathbf{B}_g - \mathbf{B}_{gg}^*\|_F \to 0$ $(p_g \to \infty)$ for $g \in \{1, 2\}$.

(A4) $\mathbf{f} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}_d)$.

2.2. Consistent estimator of ρV and its sampling distribution

The sample counterpart of ρV_{12} is obtained as

$$RV_{12} = \frac{\|\mathbf{S}_{12}\|_F^2}{\|\mathbf{S}_{11}\|_F \|\mathbf{S}_{22}\|_F},$$

where the sample covariance matrix of $\mathbf{x}_g$ and the cross-sample covariance matrix of $\mathbf{x}_1$ and $\mathbf{x}_2$ are constructed as

$$\forall g \in \{1, 2\}, \ \mathbf{S}_{gg} = \frac{1}{n-1}\sum_{i=1}^n (\mathbf{x}_{gi} - \overline{\mathbf{x}}_g)(\mathbf{x}_{gi} - \overline{\mathbf{x}}_g)^\top,$$

$$\mathbf{S}_{12} = \frac{1}{n-1}\sum_{i=1}^n (\mathbf{x}_{1i} - \overline{\mathbf{x}}_1)(\mathbf{x}_{2i} - \overline{\mathbf{x}}_2)^\top, \ \mathbf{S}_{21} = \mathbf{S}_{12}^\top$$

with $\overline{\mathbf{x}}_g = n^{-1}\sum_{i=1}^n \mathbf{x}_{gi}$ for $g \in \{1,2\}$. RV_{12} is a consistent estimator of ρV_{12} when $n \to \infty$ and p are fixed; however, it is not a consistent estimator of ρV_{12} when $n \to \infty$ and $p \to \infty$. Therefore, we define the estimator of ρV_{12} with a high-dimensionality adjustment as

$$MRV_{12} = \frac{\widehat{\|\mathbf{\Sigma}_{12}\|_F^2}}{\widehat{\|\mathbf{\Sigma}_{11}\|_F}\widehat{\|\mathbf{\Sigma}_{22}\|_F}}.$$

Here, for $g \in \{1,2\}$,

$$\widehat{\|\mathbf{\Sigma}_{gh}\|_F^2} = \frac{n-1}{n(n-2)(n-3)}[(n-1)(n-2)\mathrm{tr}(\mathbf{S}_{gh}\mathbf{S}_{hg}) + \mathrm{tr}(\mathbf{S}_{gg})\mathrm{tr}(\mathbf{S}_{hh}) - nK_{gh}],$$

where

$$K_{gh} = \frac{1}{n-1}\sum_{i=1}^n \|\mathbf{x}_{gi} - \overline{\mathbf{x}}_g\|^2\|\mathbf{x}_{hi} - \overline{\mathbf{x}}_h\|^2,$$

is an unbiased estimator of $\|\mathbf{\Sigma}_{gh}\|_F^2$ derived by [5].

Theorem 1. *Under (A1)–(A3),* $MRV_{12} = \rho V_{12} + o_p(1)$ *as* $n, p_1, p_2 \to \infty$.

To construct a hypothesis test (1), we consider the null distribution of MRV_{12}.

Theorem 2. *Suppose the null hypothesis* $\mathcal{H}$ *in (1) is true. Under (A1)–(A4),*

$$nMRV_{12} + \frac{\mathrm{tr}(\mathbf{\Lambda}_1)\mathrm{tr}(\mathbf{\Lambda}_2)}{\sqrt{\mathrm{tr}(\mathbf{\Lambda}_1^2)\mathrm{tr}(\mathbf{\Lambda}_2^2)}} \rightsquigarrow \sum_{i=1}^{d_1}\sum_{j=1}^{d_2} \frac{\lambda_{1i}\lambda_{2j}}{\sqrt{\mathrm{tr}(\mathbf{\Lambda}_1^2)\mathrm{tr}(\mathbf{\Lambda}_2^2)}}\chi_{ij}^2 \quad (n, p_1, p_2 \to \infty), \tag{4}$$

where $\chi_{11}^2, \ldots, \chi_{1d_1}^2, \chi_{21}^2, \ldots, \chi_{2d_2}^2$ *are mutually independent chi-squared distributed random variables with one degree of freedom,* $\mathbf{\Lambda}_1 = \mathrm{diag}(\lambda_{11}, \ldots, \lambda_{1d_1})$ *is* $d_1 \times d_1$ *diagonal matrix whose diagonal components are the nonzero eigenvalues of* $\mathbf{B}_{11}^*$*, and* $\mathbf{\Lambda}_2 = \mathrm{diag}(\lambda_{21}, \ldots, \lambda_{2d_2})$ *is* $d_2 \times d_2$*-diagonal matrix whose diagonal components are the nonzero eigenvalues of* $\mathbf{B}_{22}^*$.

2.3. Test procedure

By estimating the unknown parameters in the random variable on the left-hand side of (4), we construct a test statistic for (1). To estimate the number of factors d_g, we focus on the criteria function originally proposed by [1]:

$$ER_g(i) = \frac{\lambda_i(\mathbf{S}_{gg})}{\lambda_{i+1}(\mathbf{S}_{gg})},$$

where $\lambda_i(\cdot)$ is the i-th largest eigenvalue and ER_g is the eigenvalue ratio. The estimator of d_g is given by the number i that minimizes $ER_g(i)$, that is,

$$\widehat{d}_g = \underset{1 \leqslant i \leqslant i_{g,\max}}{\arg\max}\, ER_g(i),$$

where $i_{g,\max}$ denotes the prespecified upper bound of i.

We further estimate the unknown parameters $\mathrm{tr}(\boldsymbol{\Lambda}_g)$ and $\mathrm{tr}(\boldsymbol{\Lambda}_g^2)$ in (4) using

$$\widehat{\mathrm{tr}(\boldsymbol{\Lambda}_g)} = \sum_{i=1}^{\widehat{d}_g} \widehat{\lambda}_{gi} \quad \text{and} \quad \widehat{\mathrm{tr}(\boldsymbol{\Lambda}_g^2)} = \sum_{i=1}^{\widehat{d}_g} \widehat{\lambda}_{gi}^2,$$

respectively. Here, $\widehat{\lambda}_{gi} = \lambda_i(\mathbf{S}_{gg})/p_g$ for $i \in \{1, 2, \ldots, \widehat{d}_g\}$ and $g \in \{1, 2\}$.

Using these estimators, we propose a test statistic, defined as

$$T = nMRV_{12} + \frac{\widehat{\mathrm{tr}(\boldsymbol{\Lambda}_1)}\widehat{\mathrm{tr}(\boldsymbol{\Lambda}_2)}}{\sqrt{\widehat{\mathrm{tr}(\boldsymbol{\Lambda}_1^2)}\widehat{\mathrm{tr}(\boldsymbol{\Lambda}_2^2)}}}.$$

Theorem 3. *Suppose the null hypothesis $\mathcal{H}$ in (1) is true. Under (A1)–(A4),*

$$T \rightsquigarrow \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \frac{\lambda_{1i}\lambda_{2j}}{\sqrt{\mathrm{tr}(\boldsymbol{\Lambda}_1^2)\mathrm{tr}(\boldsymbol{\Lambda}_2^2)}} \chi_{ij}^2 \quad (n, p_1, p_2 \to \infty).$$

Based on the results of Theorem 3, we provide an approximate test for (1). The following are four steps of the test procedure.

1. We draw n observations from the population and calculate $\widehat{d}_g$, $\widehat{\lambda}_{gi}$ for $i \in \{1, \ldots, \widehat{d}_g\}$, $\widehat{\mathrm{tr}(\boldsymbol{\Lambda}_g)}$, and $\widehat{\mathrm{tr}(\boldsymbol{\Lambda}_g^2)}$ for $g \in \{1, 2\}$. Using these estimators, we construct T.
2. We further draw a sample of $\widehat{d}_1 \times \widehat{d}_2$ independently and χ_{ij}^2-distributed random variables to obtain

$$\widetilde{T} = \sum_{i=1}^{\widehat{d}_1} \sum_{j=1}^{\widehat{d}_2} \frac{\widehat{\lambda}_{1i}\widehat{\lambda}_{2j}}{\sqrt{\widehat{\mathrm{tr}(\boldsymbol{\Lambda}_1^2)}\widehat{\mathrm{tr}(\boldsymbol{\Lambda}_2^2)}}} \chi_{ij}^2.$$

3. We then repeat step 2 until we obtain a Monte Carlo estimate of the distribution for the random variable $\widetilde{T}$ and its $(1-\alpha)$-quantile $\hat{t}_\alpha$.
4. We further realized an approximate test with the nominal size α as follows:

$$\text{Reject } \mathcal{H} \overset{\text{def}}{\Longleftrightarrow} T > \widehat{t}_\alpha. \tag{5}$$

2.4. Aspects of power

To examine the power of test (5), we consider the following local alternatives:

$\mathcal{A}_L$: Let η be a constant greater than or equal to 1/2. There exists a $d \times d$ matrix $\boldsymbol{\Xi}$ such that all diagonal elements are 0 and at least one off-diagonal element is not 0 such that the following condition is met:

$$\left\| \frac{n^\eta}{p_1 p_2} \mathbf{B}_1^\top \mathbf{B}_1 \mathbf{B}_2^\top \mathbf{B}_2 - \boldsymbol{\Xi} \right\|_F \to 0 \quad (n, p_1, p_2 \to \infty).$$

Furthermore, there exists a positive real number Δ such that the following condition is met:

$$\frac{n^{2\eta}}{p_1p_2}\|\mathbf{\Sigma}_{12}\|_F^2 = \frac{n^{2\eta}}{p_1p_2}\mathrm{tr}(\mathbf{B}_1^\top\mathbf{B}_1\mathbf{B}_2^\top\mathbf{B}_2) \to \Delta \quad (n, p_1, p_2 \to \infty).$$

Theorem 4. *Under the local alternatives $\mathcal{A}_L$ and (A1)–(A4),*

$$nMRV_{12} + \frac{\mathrm{tr}(\mathbf{\Lambda}_1)\mathrm{tr}(\mathbf{\Lambda}_2)}{\|\mathbf{\Lambda}_1\|_F\|\mathbf{\Lambda}_2\|_F} \rightsquigarrow \begin{cases} \Delta/(\|\mathbf{\Lambda}_1\|_F\|\mathbf{\Lambda}_2\|_F) + \mathbf{z}^\top\mathbf{C}^*\mathbf{z} + \mathrm{c}^{*\top}\mathbf{z} & \eta = 1/2, \\ \mathbf{z}^\top\mathbf{C}^*\mathbf{z} & \eta > 1/2, \end{cases}$$

where $\mathbf{z}$ has a d^2-variate normal distribution with a mean vector $\mathbf{0}$ and covariance matrix $\mathbf{I}_{d^2} + \mathbf{K}_{d^2}$ and

$$\mathbf{C}^* = \frac{1}{\|\mathbf{\Lambda}_1\|_F\|\mathbf{\Lambda}_2\|_F}(\mathbf{B}_{11}^* \otimes \mathbf{B}_{22}^*), \ \ \mathrm{c}^* = \frac{1}{\|\mathbf{\Lambda}_1\|_F\|\mathbf{\Lambda}_2\|_F}\mathrm{vec}(\mathbf{\Xi} + \mathbf{\Xi}^\top).$$

Here, $\mathbf{K}_{d^2}$ denotes the commutation matrix.

Applying the theorem, we obtain the following corollary of the asymptotic power under local alternative $\mathcal{A}_L$.

Corollary 1. *Under (A1)–(A4), the asymptotic power function is*

$$\Pr(T > \widehat{t}_\alpha|\mathcal{A}_L) = \begin{cases} G\{t_\alpha - \Delta/(\|\mathbf{\Lambda}_1\|_F\|\mathbf{\Lambda}_2\|_F)\} + o(1) & \eta = 1/2, \\ \alpha + o(1) & \eta > 1/2, \end{cases}$$

where $G(\cdot)$ denotes the cumulative distribution function of $\mathbf{z}^\top\mathbf{C}^*\mathbf{z} + \mathrm{c}^{*\top}\mathbf{z}$.

3. Numerical studies

We examine the size and power of test (5) in a finite sample and dimension by Monte Carlo simulations.

References

[**1**] Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica* **81**, 1203–1227.

[**2**] Escoufier, Y. (1973). Le Traitement des variables vectorielles. *Biometrics* **29**, 751–760.

[**3**] Hyodo, M., Nishiyama, T., and Pavlenko, T. (2020). Testing for independence of high-dimensional variables:ρV-coefficient based approach. *J. Multivariate Anal.* **178**, 104627.

[**4**] Josse, J. and Holmes, S. (2016). Measuring multivariate association and beyond. *Statistics Surveys* **10**, 132–167.

[**5**] Yamada, Y., Hyodo, M., and Nishiyama, T. (2017). Testing block-diagonal covariance structure for high-dimensional data under non-normality. *J. Multivariate Anal.* **155**, 305–316.