On a general linear hypothesis testing problem for latent factor models in high dimensions

Takahiro Nishiyama[a] and Masashi Hyodo[b]

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

1. Introduction

Let $\mathbf{x}_{gi} = (x_{gi1}, \ldots, x_{gip})^\top \sim \mathcal{F}_g$ be iid p-dimensional random vectors collected from the ith subject in the gth population, where $\mathcal{F}_g$ denotes the distribution function for gth population, $i \in \{1, \ldots, n_g\}$, $g \in \{1, \ldots, k\}$. A factor model assumes that for each $g \in \{1, \ldots, k\}$, the observable vector $\mathbf{x}_{gi}$ is decomposable into a latent factor and an idiosyncratic component as follows:

$$\mathbf{x}_{gi} = \boldsymbol{\mu}_g + \mathbf{F}_g \mathbf{z}_{gi} + \boldsymbol{\Psi}_g^{1/2} \boldsymbol{\epsilon}_{gi}, \tag{1}$$

where $\boldsymbol{\mu}_g \in \mathbb{R}^p$ is a deterministic intercept vector, $\mathbf{z}_{gi} = (\mathrm{z}_{gi1}, \ldots, \mathrm{z}_{gid_g})^\top$ is a d_g-dimensional latent factor vector, and $\boldsymbol{\epsilon}_{gi} = (\epsilon_{gi1}, \ldots, \epsilon_{gip})^\top$ is a p-dimensional error vector which is uncorrelated with the latent factor. In what follows, we assume that $d_g \in \mathbb{N}$ is a fixed number. Further, $\mathbf{F}_g = (\mathbf{f}_{g1}, \ldots, \mathbf{f}_{gp})^\top$ denotes a loading matrix where for each $j \in \{1, \ldots, p\}$, $\mathbf{f}_{gj} = (f_{gj1}, \ldots, f_{gjd_g})^\top \in \mathbb{R}^{d_g}$ is a non-random vector, and $\boldsymbol{\Psi}_g = \mathrm{diag}(\psi_{g1}, \ldots, \psi_{gp})$ is a non-random $p \times p$ diagonal matrix whose elements are $\psi_{g1} > 0, \ldots, \psi_{gp} > 0$. For the latent vector $\mathbf{z}_{gi}$ and error vector $\boldsymbol{\epsilon}_{gi}$, we further assume that $\mathrm{z}_{gi\ell}$ are iid with $\mathrm{E}(\mathrm{z}_{gi\ell}) = 0$, $\mathrm{E}(\mathrm{z}_{gi\ell}^2) = 1$ and $\mathrm{E}(\mathrm{z}_{gi\ell}^4) = \kappa_{\mathrm{z}_g} < \infty$, and ϵ_{gij} are iid with $\mathrm{E}(\epsilon_{gij}) = 0$, $\mathrm{E}(\epsilon_{gij}^2) = 1$ and $\mathrm{E}(\epsilon_{gij}^4) = \kappa_{\epsilon_g} < \infty$ for $g \in \{1, \ldots, k\}$, $i \in \{1, \ldots, n_g\}$, $j \in \{1, \ldots, p\}$ and $\ell \in \{1, \ldots, d_g\}$. Structural assumptions of the model (1) imply that

$$\mathrm{E}(\mathbf{x}_{gi}) = \boldsymbol{\mu}_g, \ \mathrm{cov}(\mathbf{x}_{gi}) = \mathbf{F}_g \mathbf{F}_g^\top + \boldsymbol{\Psi}_g := \boldsymbol{\Sigma}_g, \tag{2}$$

where $\boldsymbol{\Sigma}_g \in \mathbb{R}_{>0}^{p \times p}$ and $\mathbb{R}_{>0}^{p \times p}$ denotes the space of real, symmetric, positive definite, $p \times p$ matrices.

By using the data generated by (1), we design a high-dimensinal test procedure for a general linear hypothesis testing (GLHT) problem:

$$\mathcal{H} : \widetilde{\mathbf{G}}\mathbf{M} = \mathbf{O}, \quad \mathcal{A} : \widetilde{\mathbf{G}}\mathbf{M} \neq \mathbf{O}, \tag{3}$$

where $\mathbf{M} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k)^\top$ is a $k \times p$ matrix and $\widetilde{\mathbf{G}}$ is a $q \times k$ known coefficient matrix with full row rank $q < k$. By setting $\widetilde{\mathbf{G}}$ to be any $(k-1) \times k$ contrast matrix, i.e., any $(k-1) \times k$ matrix with linearly independent rows and zero row sums, the GLHT problem (3) reduces to the one-way MANOVA problem:

$$\mathcal{H} : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_k, \quad \mathcal{A} :\neq \mathcal{H}. \tag{4}$$

Also, various post hoc and contrast tests can be written in the form of (3).

2. Methodology and main results

2.1. Asymptotic set up

We shall now formalize the asymptotic viewpoint which we adopt for the latent factor model of our interest. Our analysis takes place in an asymptotic setting where both the number of features p as well as the sample sizes n_g go to infinity. In the model (1), we choose the loading matrix $\mathbf{F}_g$ and the noise covariance matrix $\mathbf{\Psi}_g$ according to n_g and p with the following assumptions.

(A1) q and k are fixed. $p = p(n) \to \infty$ as a function of $n = n_1 + \cdots + n_k$ such that p tents to infinity along with $n \to \infty$, $n/p \to \eta \in (0, \infty)$, $n_g \to \infty$ in such a way that $n_g/n \to \gamma_g \in (0, \infty)$, $a_{gg} \to \tau_{gg} \in (0, \infty)$, and $a_{gh} \to \tau_{gh} \in (-\infty, \infty)$ which satisfies $\tau_{gh} = \tau_{hg}$ for $g \neq h$. Here, a_{gh} is element of the matrix $\mathbf{A} = \mathbf{D}^{1/2}\mathbf{H}\mathbf{D}^{1/2}$. Note that $\mathbf{A}$ is an idempotent matrix, i.e., $\mathbf{A} = \mathbf{A}^\top$, $\mathbf{A}^2 = \mathbf{A}$, and $\mathrm{tr}(\mathbf{A}) = q$.

(A2) Let $\psi_{g\max} = \max\{\psi_{g1}, \ldots, \psi_{gp}\}$ for $g \in \{1, \ldots, k\}$. Then, $\psi_{g\max}/p^{1/2} \to 0$ and $(1/p)\mathbf{F}_g^\top \mathbf{F}_h \to \mathbf{U}_{gh}$ for $h \in \{1, \ldots, k\}$ as $p \to \infty$, where $\mathbf{U}_{gg}$ is $d_g \times d_g$ positive definite matrix and $\mathbf{U}_{gh}$ is $d_g \times d_h$ real matrix which satisfies $\mathbf{U}_{gh}^\top = \mathbf{U}_{hg}$.

2.2. Asymptotic distribution theory

From Zhang et al. (2017) and Zhang et al. (2023), we re-write (3) into the following equivalent form:

$$\mathcal{H} : \mathbf{C}\boldsymbol{\mu} = \mathbf{0}, \quad \mathcal{A} : \mathbf{C}\boldsymbol{\mu} \neq \mathbf{0}, \tag{5}$$

where $\mathbf{C} = \mathbf{G} \otimes \mathbf{I}_p$ ($qp \times kp$ matrix), $\mathbf{G} = (\widetilde{\mathbf{G}}\mathbf{D}\widetilde{\mathbf{G}}^\top)^{-1/2}\widetilde{\mathbf{G}}$ with $\mathbf{D} = \mathrm{diag}(1/n_1, \cdots, 1/n_k)$ and $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^\top, \ldots, \boldsymbol{\mu}_k^\top)^\top$.

For testing (5), an L^2-norm based test statistic can be constructed as follows:

$$T = \|\mathbf{C}\widehat{\boldsymbol{\mu}}\|^2 = \widehat{\boldsymbol{\mu}}^\top(\mathbf{H} \otimes \mathbf{I}_p)\widehat{\boldsymbol{\mu}}, \tag{6}$$

where $\mathbf{H} = \mathbf{G}^\top\mathbf{G}$ and $\widehat{\boldsymbol{\mu}} = (\overline{\mathbf{x}}_1^\top, \ldots, \overline{\mathbf{x}}_k^\top)^\top$. Here, $\overline{\mathbf{x}}_g = n_g^{-1}\sum_{i=1}^{n_g}\mathbf{x}_{gi}$ for $g \in \{1, \ldots, k\}$.

Theorem 1. *Suppose that the null hypothesis $\mathcal{H}$ from (3) is true. For a latent factor model (1) satisfying conditions (A1) and (A2), random variable $\{T - \sum_{g=1}^k a_{gg}\mathrm{tr}(\mathbf{\Psi}_g)\}/p$ is asymptotically distributed as $\sum_{\ell=1}^d \lambda_\ell(\mathbf{V})\chi_\ell^2(1)$ as $\min\{n_1, \ldots, n_k\} \to \infty$, where $\chi_1^2(1), \ldots, \chi_d^2(1)$ are mutually independent, chi-square distributed random variables with 1 degree of freedom and $\lambda_\ell(\mathbf{V})$ is the ℓth largest eigenvalue of matrix $\mathbf{V}$. Here, $d = d_1 + \cdots + d_k$,*

$$\mathbf{V} = \begin{pmatrix} \tau_{11}\mathbf{U}_{11} & \cdots & \tau_{1k}\mathbf{U}_{1k} \\ \vdots & \vdots & \vdots \\ \tau_{k1}\mathbf{U}_{k1} & \cdots & \tau_{kk}\mathbf{U}_{kk} \end{pmatrix}.$$

2.3. Test procedure

In practice, the consistency is expected to hold with unknown parameters replaced by their estimators. To estimate the number of factors d_g, we focus on the criteria function which is proposed by Ahn and Horenstein (2013)

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

where ER_g refers to eigenvalue ratio and $\mathbf{S}_g = 1/(n_g - 1)\sum_{i=1}^{n_g}(\mathbf{x} - \overline{\mathbf{x}}_g)(\mathbf{x} - \overline{\mathbf{x}}_g)^\top$. The estimator of d_g is i which minimizes the $ER_g(i)$, that is

$$\widehat{d}_g = \max_{1 \leqslant i \leqslant i_{g,\max}} ER_g(i). \tag{7}$$

Suppose that the factor model (1) satisfies conditions (A1)-(A2). Then, there exists $c_g \in (0, 1]$ such that $\mathbb{P}(\widehat{d}_g = d_g) \to 1$ as $\min\{p, n_g\} \to \infty$, for any $i_{g,\max} \in (d_g, \lfloor c_g \min\{p, n_g\} \rfloor - d_g - 1]$, where $\lfloor \cdot \rfloor$ denotes the floor function.

We also estimate the unknown parameters $\mathrm{tr}(\mathbf{\Psi}_g)$ in T,

$$\widehat{\mathrm{tr}(\mathbf{\Psi}_g)} = \mathrm{tr}(\mathbf{S}_g) - \sum_{i=1}^{\widehat{d}_g} \lambda_i(\mathbf{S}_g)$$

respectively. By using these estimators, we propose the test statistic defined as

$$T_{nh} = \frac{1}{p}\left\{ T - \sum_{g=1}^{k} a_{gg} \widehat{\mathrm{tr}(\mathbf{\Psi}_g)} \right\}$$

Theorem 2. *Suppose that the null hypothesis $\mathcal{H}$ from (3) is true. For a latent factor model (1) satisfying conditions (A1) and (A2), random variable T_{nh} is asymptotically distributed as $\sum_{\ell=1}^{d} \lambda_\ell(\mathbf{V})\chi^2_\ell(1)$.*

Now, on the basis the results of Theorems 1 and 2 we furnish approximation test of significance of $\mathcal{H}$. The following are four steps of the test procedure.

1. For each $g \in \{1, \ldots, k\}$, draw n_g observations from population G_g and calculate $\widehat{d}_g$, T_{nh}.

2. Let $n_0 = \min\{n_1, \ldots, n_k\}$. For each $i \in \{1, \ldots, n_0\}$, let $\mathbf{w}_i = (\mathbf{G}\mathbf{D}^{1/2} \otimes \mathbf{I}_p)\mathbf{x}_i$ where $\mathbf{x}_i = (\mathbf{x}_{1i}^\top, \ldots, \mathbf{x}_{ki}^\top)^\top$ and calclate the covariance matrix as

$$\mathbf{S}_w = \frac{1}{n_0 - 1}\sum_{i=1}^{n_0}(\mathbf{w}_i - \overline{\mathbf{w}})(\mathbf{w}_i - \overline{\mathbf{w}})^\top.$$

   Using $\mathbf{S}_w$, estimate $\lambda_\ell(\mathbf{V})$ as $\widehat{\lambda_\ell(\mathbf{V})} = \lambda_\ell(\mathbf{S}_w)/p$ for $\ell \in \{1, \ldots, \sum_{g=1}^{k} \widehat{d}_g\}$.

3. Let $\widehat{d} = \widehat{d}_1 + \cdots + \widehat{d}_k$. By utilizing Monte Carlo simulation, we calculate the estimator $\widehat{t}_\alpha$ of true $(1 - \alpha)$-quantile of the distribution $\sum_{\ell=1}^{\widehat{d}} \widehat{\lambda_\ell(\mathbf{V})}\chi^2_\ell(1)$.

4. We can obtain an approximate test with the nominal size α as follows:

$$\text{Reject } \mathcal{H} \overset{\text{def}}{\Longleftrightarrow} T_{nh} > \widehat{t}_{\alpha}. \tag{8}$$

2.4. Aspects of power

Next, we examine the asymptotic power of the test (8). The following theorem gives the distribution of $T_{nh} - \|\mathbf{C}\boldsymbol{\mu}\|^2/p$ under alternative hypotheses for examining asymptotic power.

Theorem 3. *For a latent factor model (1) satisfying conditions (A1), (A2), and* $\|\mathbf{C}\boldsymbol{\mu}\| = O(p)$*, random variable* $T_{nh} - \|\mathbf{C}\boldsymbol{\mu}\|^2/p$ *is asymptotically distributed as*

$$\tilde{\mathbf{z}}_0^{\top}\mathbf{V}\tilde{\mathbf{z}}_0 + 2(\mathbf{C}\boldsymbol{\mu})^{\top}(\mathbf{G}\mathbf{D}^{1/2} \otimes \mathbf{I}_p)\mathbf{F}\tilde{\mathbf{z}}_0/p, \tag{9}$$

where $\tilde{\mathbf{z}}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$.

Using Theorem 3, we can derive the following corollary for the power of test (8).

Corollary 1. *Let* $\|\mathbf{C}\boldsymbol{\mu}\|_F^2 \asymp p^{\delta}$*. For a latent factor model (1) satisfying conditions (A1) and (A2),*

$$\textit{power} = \begin{cases} \alpha + o(1) & \textit{under} \;\; \delta < 1, \\ 1 - G(t_{\alpha} - \|\mathbf{C}\boldsymbol{\mu}\|_F^2/p) + o(1) & \textit{under} \;\; \delta = 1, \\ 1 + o(1) & \textit{under} \;\; \delta > 1. \end{cases}$$

Here, $G(\cdot)$ *is the cdf of (9) and* t_{α} *is true* $(1-\alpha)$*-quantile of the distribution* $\sum_{\ell=1}^{d} \lambda_{\ell}(\mathbf{V})\chi_{\ell}^2(1)$.

3. Numerical studies

Assuming a MANOVA model for comparing 3 population mean vectors, we compare, through simulations, the performance of the proposed test and existing procedures in terms of size control and power.

References

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

[2] Zhang, J.-T., Guo, J., Zhou, B., 2017. Linear hypothesis testing in high-dimensional one-way MANOVA. *J. Multivar. Anal.*, **155**, 200–216.

[3] Zhang, J.-T., Zhou, B., Guo, J., 2022. Linear hypothesis testing in high-dimensional heteroscedastic one-way MANOVA: A normal reference L^2-norm based test. *J. Multivar. Anal.*, **187**, 104816.