2021年6月18日　統計数理研究所　オープンハウス

Higher-order approximation of the distribution of test statistics for high-dimensional time-series ANOVA models

Hideaki NAGAHATA (Risk Analysis Research Center, Institute of Statistical Mathematics)

Classical ANOVA works well for high dimensional time series. For example, this method can be applied for radioactive data.

Analysis of variance (ANOVA) is tailored for independent observations. Recently, there has been considerable demand for ANOVA of high-dimensional and dependent observations in many fields. For example, it is important to analyze differences among industry averages of financial data. However, ANOVA for these types of observations has been inadequately developed. In this paper, we thus present a study of ANOVA for high-dimensional and dependent observations. Specifically, we present the asymptotics of classical test statistics proposed for independent observations and provide a sufficient condition for them to be asymptotically normal. Numerical examples for simulated and radioactive data are presented as applications of these results.

1 Theoretical results

1.1 Setting

Let p vetor-valued series $\boldsymbol{X}_{i1}, \dots \boldsymbol{X}_{in_i}$ $(p \to \infty)$ be generated from

$$\boldsymbol{X}_{it} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_{it},\ t = 1, \dots, n_i,\ i = 1, \dots, q,$$

where

- $\boldsymbol{\epsilon}_i \equiv \{\boldsymbol{\epsilon}_{it};\ t = 1, \dots, n_i\},\ i = 1, \dots, q$, are stationary with mean $\mathbf{0}$, autocovariance matrix $\boldsymbol{\Gamma}(\cdot)$ and spectral density matrix $\boldsymbol{f}(\lambda)$,
- $\{\boldsymbol{\epsilon}_{it};\ t = 1, \dots, n_i\},\ i = 1, \dots, q$, are mutually independent.

Consider the problem of testing

$$H : \boldsymbol{\alpha}_1 = \cdots = \boldsymbol{\alpha}_q.$$

Assumption 1 (high dimensional large sample setting)**.**

$$\frac{p^{3/2}}{\sqrt{n}} \to 0 \ as\ n, p \to \infty,$$

$$\frac{n_i}{n} \to \rho_i > 0 \ as\ n \to \infty,\ n = \sum_{i=1}^{q} n_i.$$

1.2 Method

- For independent observations, the following Lawley-Hotelling test (1), likelihood ratio test (2), and Bartlett-Nanda-Pillai test (3) have been proposed:

$$LH \equiv n\mathrm{tr}\{\hat{\mathcal{S}}_H \hat{\mathcal{S}}_E^{-1}\}, \quad (1)$$
$$LR \equiv -n \log\{|\hat{\mathcal{S}}_E|/|\hat{\mathcal{S}}_E + \hat{\mathcal{S}}_H|\}, \quad (2)$$
$$BNP \equiv n\mathrm{tr}\hat{\mathcal{S}}_H(\hat{\mathcal{S}}_E + \hat{\mathcal{S}}_H)^{-1}, \quad (3)$$

where

$$\hat{\mathcal{S}}_H \equiv \sum_{i=1}^{q} n_i(\hat{\boldsymbol{X}}_{i\cdot} - \hat{\boldsymbol{X}}_{\cdot\cdot})(\hat{\boldsymbol{X}}_{i\cdot} - \hat{\boldsymbol{X}}_{\cdot\cdot})',$$

$$\hat{\mathcal{S}}_E \equiv \sum_{i=1}^{q} \sum_{t=1}^{n_i} (\boldsymbol{X}_{it} - \hat{\boldsymbol{X}}_{i\cdot})(\boldsymbol{X}_{it} - \hat{\boldsymbol{X}}_{i\cdot})'.$$

- We can derive the stochastic expansion of the standardized versions T_1, T_2, T_3 of three tests LH, LR, BNP respectively;

$$T_1 \equiv \frac{1}{\sqrt{2(q-1)}} \left\{ \frac{1}{\sqrt{p}} LH - \sqrt{p}(q-1) \right\},$$
$$T_2 \equiv \frac{1}{\sqrt{2(q-1)}} \left\{ \frac{1}{\sqrt{p}} LR - \sqrt{p}(q-1) \right\},$$
$$T_3 \equiv \frac{1}{\sqrt{2(q-1)}} \left\{ \frac{1}{\sqrt{p}} BNP - \sqrt{p}(q-1) \right\}.$$

1.3 Results

Assumption 2 (Brillinger condition)**.** *Given a p-vector stationary process* $\boldsymbol{\epsilon}_{it} = (\epsilon_{it}^{(1)}, \cdots, \epsilon_{it}^{(p)})'$ *for each* $k = 2, 3, \cdots,$ *and* $j = 1, \cdots, k-1$, *there exists an* $m > 0$ *with*

$$\sum_{t_1, \dots, t_{k-1} = -\infty}^{\infty} \{1 + |t_j|\}^m |c_{a_1, \cdots, a_k}(t_1, \cdots, t_{k-1})| < \infty$$

uniformly for $a_1, \cdots, a_k$, *where* $c_{a_1, \cdots, a_k}(t_1, \cdots, t_{k-1}) = \mathrm{cum}\{\epsilon_{it_1}^{(a_1)}, \cdots, \epsilon_{it_k}^{(a_k)}\}$.

Assumption 3 (Uncorrelated disturbance)**.**

$$\boldsymbol{\Gamma}(j) = \mathbf{0} \ for\ all\ j \neq 0.$$

Remark 1. *Assumption 3 is not severe because vector GARCH model (very practical nonlinear time series model) satisfies it.*

Theorem 1. *Suppose Assumptions 1-3. Then, under the null hypothesis H, we have the following Edgeworth expansions:*

$$P(T_i < z) = \Phi(z) - \phi(z)\left\{ p^{-1/2} \cdot \frac{c_3}{6}(z^2 - 1) + p^{-1} \cdot \frac{c_4}{24}(z^3 - 3z) \right\} + o\left(p^{-1}\right), \quad (i = 1, 2, 3)$$

where

$$\Phi(z) = \int_{-\infty}^{z} \phi(y)dy, \quad \phi(y) = (2\pi)^{-1/2} exp\left(-\frac{y^2}{2}\right),$$

and

$$c_3 = \left(\frac{2}{q-1}\right)^{3/2} \left\{ q - 3 + 3\sum_{i=1}^{q} \left(\frac{n_i}{n}\right)^2 - \sum_{i=1}^{q} \left(\frac{n_i}{n}\right)^3 \right\},$$
$$c_4 = \left(\frac{2}{q-1}\right)^{2} \left\{ q - 4 + 6\sum_{i=1}^{q} \left(\frac{n_i}{n}\right)^2 - 4\sum_{i=1}^{q} \left(\frac{n_i}{n}\right)^3 - \sum_{i=1}^{q} \left(\frac{n_i}{n}\right)^4 \right\}.$$

2 Simulation results

Calculate $\hat{F}_{i,n}(z),\ i = 1, 2, 3$, which is the empirical distribution of $\{T_i^{(1)}, \dots, T_i^{(1000)};\ i = 1, 2, 3\}$.

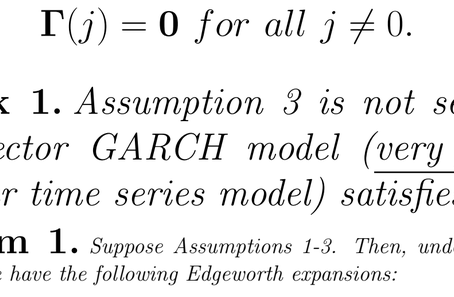


Figure 1: Plot of $\{|\hat{F}_{1,n}(z) - \Phi(z)| - |\hat{F}_{1,n}(z) - P(T_1 < z)|\}$ by dotted and thick lines, respectively.

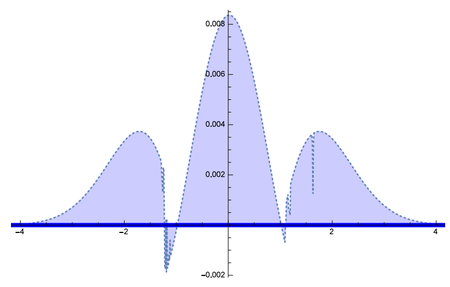


Figure 2: Plot of $\{|\hat{F}_{2,n}(z) - \Phi(z)| - |\hat{F}_{2,n}(z) - P(T_2 < z)|\}$ by dotted and thick lines, respectively.

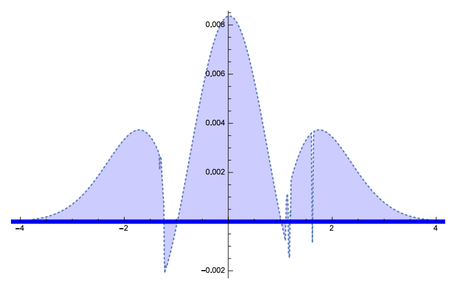
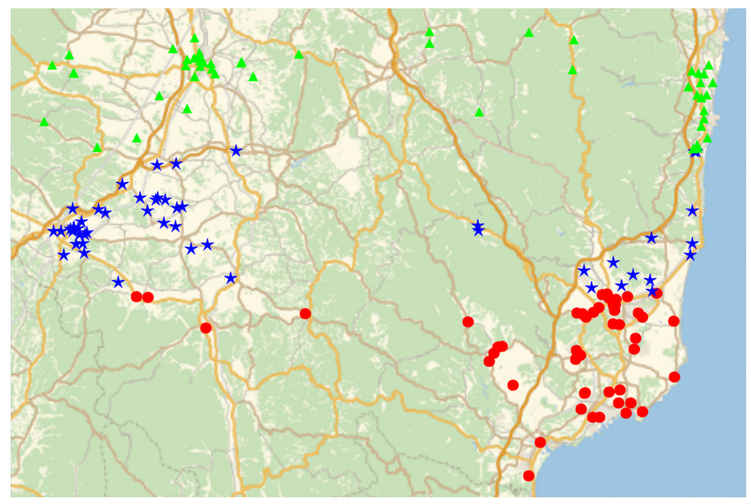


Figure 3: Plot of $\{|\hat{F}_{3,n}(z) - \Phi(z)| - |\hat{F}_{3,n}(z) - P(T_3 < z)|\}$ by dotted and thick lines, respectively.

3 Application to radioactive data



- We apply T_1, T_2, T_3 to the radioactive data of Fukushima.
- Data: This data set consists of three groups with 50 dimensions and about 8000 cell lines.
  - 3 groups, (i) Green area, (ii) Blue area, and (iii) Red area.
  - Very low autocorrelation;

→ All of the tests reject hypothesis H, and their P-values are all around 0.