2019年6月5日　統計数理研究所　オープンハウス

Analysis of variance for high dimensional time series

Hideaki Nagahata* and Masanobu Taniguchi　(Risk Analysis Research Center, Institute of Statistical Mathematics(*) / Waseda University)

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)**.**

$$\frac{p}{\sqrt{n}} \to 0 \ as\ n, p \to \infty,$$

$$\frac{n_i}{n} \to \rho_i > 0 \ as\ n \to \infty.$$

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,*

$$T_i \xrightarrow{d} N(0, 1), \quad i = 1, 2, 3.$$

2 Simulation results

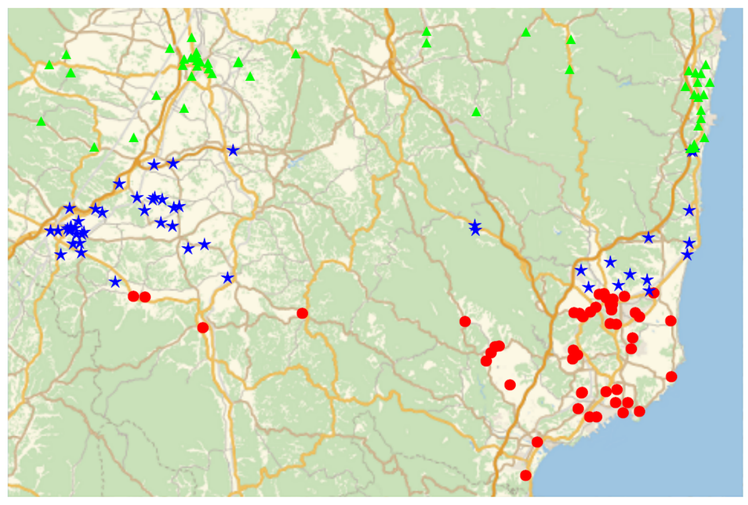
Sample size of each group	Test Statistic	Significance Level 10%	5%	1%
50	T_1 (LH)	0.906	0.879	0.781
	T_2 (LR)	0.615	0.493	0.313
	T_3 (BNP)	0.089	0.037	0.010
100	T_1 (LH)	0.564	0.460	0.286
	T_2 (LR)	0.323	0.212	0.098
	T_3 (BNP)	0.107	0.060	0.015
500	T_1 (LH)	0.161	0.100	0.032
	T_2 (LR)	0.134	0.082	0.022
	T_3 (BNP)	0.101	0.055	0.015
2500	T_1 (LH)	0.114	0.062	0.015
	T_2 (LR)	0.108	0.058	0.014
	T_3 (BNP)	0.100	0.057	0.014
7500	T_1 (LH)	0.109	0.058	0.012
	T_2 (LR)	0.107	0.055	0.012
	T_3 (BNP)	0.105	0.054	0.012

Table 1: Rejection rate of the test statistics.

Sample size of each group	Test Statistic	Significance Level 10%	5%	1%
50	T_1 (LH)	0.960	0.939	0.890
	T_2 (LR)	0.768	0.684	0.520
	T_3 (BNP)	0.200	0.108	0.020
100	T_1 (LH)	0.844	0.770	0.606
	T_2 (LR)	0.636	0.518	0.346
	T_3 (BNP)	0.352	0.230	0.086
500	T_1 (LH)	0.978	0.959	0.907
	T_2 (LR)	0.972	0.950	0.864
	T_3 (BNP)	0.958	0.934	0.821
2500	T_1 (LH)	1.000	1.000	1.000
	T_2 (LR)	1.000	1.000	1.000
	T_3 (BNP)	1.000	1.000	1.000
7500	T_1 (LH)	1.000	1.000	1.000
	T_2 (LR)	1.000	1.000	1.000
	T_3 (BNP)	1.000	1.000	1.000

Table 2: Powers of the test statistics under the alternative hypothesis (ii) $\boldsymbol{\alpha}_1 = (-0.01, \cdots, -0.01)'$, $\boldsymbol{\alpha}_2 = \mathbf{0}$, $\boldsymbol{\alpha}_3 = (0.01, \cdots, 0.01)'$.

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.