Discriminant and cluster analysis of possibly high-dimensional time series data by a class of disparities

Yan Liu, Hideaki Nagahata*, Hirotaka Uchiyama and Masanobu Taniguchi (Waseda University /Project Researcher, Risk Analysis Research Center, Institute of Statistical Mathematics(*))

For possibly high-dimensional time series data, a basic discriminant statistic has a goodness, and this can be applied to a classification of companies.

Discriminant and cluster analysis of high-dimensional time series data have been an urgent need in more and more academic fields. For possibly high-dimensional and stationary time series data, we show the consistency of classifier under suitable conditions. Also, simulation studies show that that works even in the case of finite observations of training samples. Finally, we conduct the cluster analysis for real financial data. We conclude that our method is suitable for the discriminant and cluster analysis of high-dimensional dependent data.

1 Discriminant analysis

1.1 Setting

$\boldsymbol{X} = \{\boldsymbol{X}(t) = (X_1(t), \dots, X_p(t))' ; t \in \mathbb{Z}\}$ ($p < \infty$ *or* $p \to \infty$) :possibly high dimensional stationary process with

- mean $\boldsymbol{\mu}$ & spectral density matrix $\boldsymbol{f}(\lambda)$.
- Categories :

$$\pi_1 : \boldsymbol{\mu} = \boldsymbol{\mu}^{(1)}, \quad \boldsymbol{f}(\lambda) = \boldsymbol{f}^{(1)}(\lambda),$$

$$\pi_2 : \boldsymbol{\mu} = \boldsymbol{\mu}^{(2)}, \quad \boldsymbol{f}(\lambda) = \boldsymbol{f}^{(2)}(\lambda).$$

Step 1 Independent training samples $\boldsymbol{X}^{(1)}$ & $\boldsymbol{X}^{(2)}$ from π_1 & π_2 with size n_1 & n_2, respectively, are available.

Step 2 We obtain samples $\boldsymbol{X}$ with size n.

$$\Downarrow$$

Classification problem : $\boldsymbol{X} \in \pi_1$ or π_2 ?

1.2 Method and result

- To classify $\boldsymbol{X}$ by

$$\boldsymbol{\Gamma}(\boldsymbol{X}) = \left(\bar{\boldsymbol{X}} - \frac{\bar{\boldsymbol{X}}^{(1)} + \bar{\boldsymbol{X}}^{(2)}}{2}\right)'(\bar{\boldsymbol{X}}^{(2)} - \bar{\boldsymbol{X}}^{(1)}),$$

where

$$\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{X}(t), \quad \bar{\boldsymbol{X}}^{(i)} = \frac{1}{n_i}\sum_{t=1}^{n_i} \boldsymbol{X}^{(i)}(t).$$

- Rule:
  $-\boldsymbol{\Gamma}(\boldsymbol{X}) < 0 \to \boldsymbol{X}$ into π_1,
  $-\boldsymbol{\Gamma}(\boldsymbol{X}) \geq 0 \to \boldsymbol{X}$ into π_2.
- Purpose: consistent classifier is $P(i|j) \to 0$ for $(i, j) = (1, 2), (2, 1)$.

Assumption 1. *There exists* $\eta \geq 0$ *such that* $c_1 p^\eta < \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 < c_2 p^\eta$.

Assumption 2. *(I)* p *is finite and* $n_1, n \to \infty$ $(\eta \geq 0)$,

(II) $p \to \infty$, *and*

(i) if $\eta > 1$, *then, both* n_1 *and* n *are finite or infinite,*

(ii) if $\eta = 1$, *then,* $n_1 \to \infty$ *and* n *is finite or infinite,*

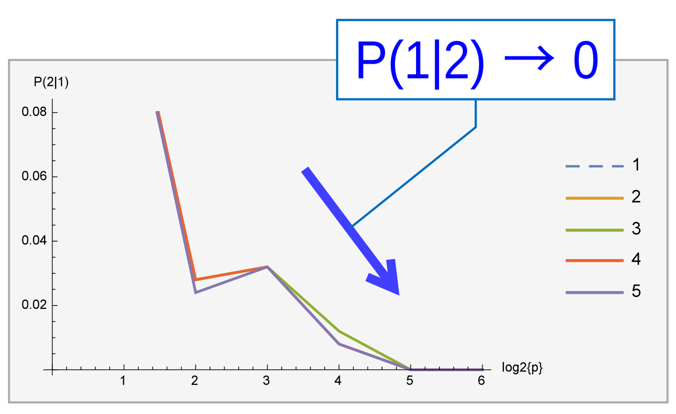
(iii) if $1/2 < \eta < 1$, *then,* $n_1 \to \infty$ *and* n *is finite or infinite, such that* $p = o(n_1^{1/(1-\eta)})$,

(iv) if $\eta = 1/2$, *then,* $n_1, n \to \infty$, *such that* $p = o(n_1^2)$,

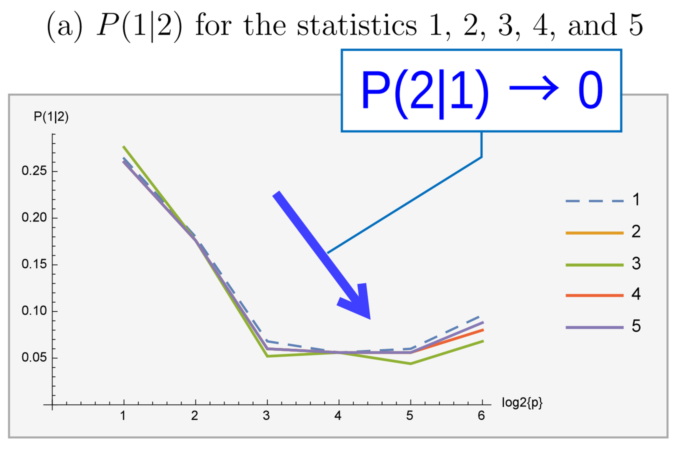
(v) if $0 \leq \eta < 1/2$, *then,* $n_1, n \to \infty$, *such that* $p = o\big((n_1^{1/(1-\eta)} n^{2/(1-2\eta)})/(n_1^{1/(1-\eta)} + n^{2/(1-2\eta)})\big)$.

Theorem 1. *Under some appropriate assumptions,* $\boldsymbol{\Gamma}(\boldsymbol{X})$ *is a consistent classifier.*

1.3 Simulation

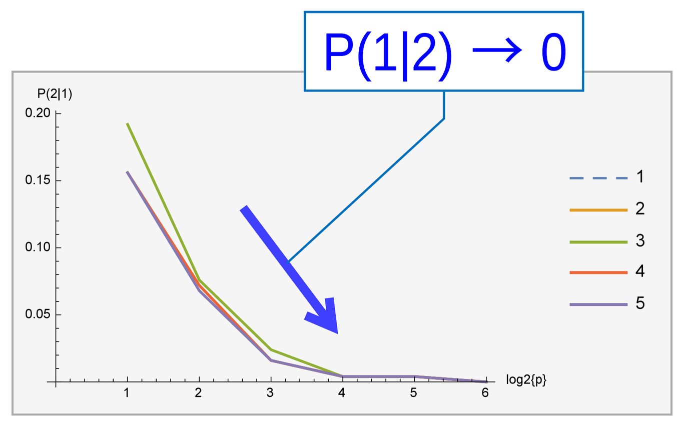


(a) $P(1|2)$ for the statistics 1, 2, 3, 4, and 5

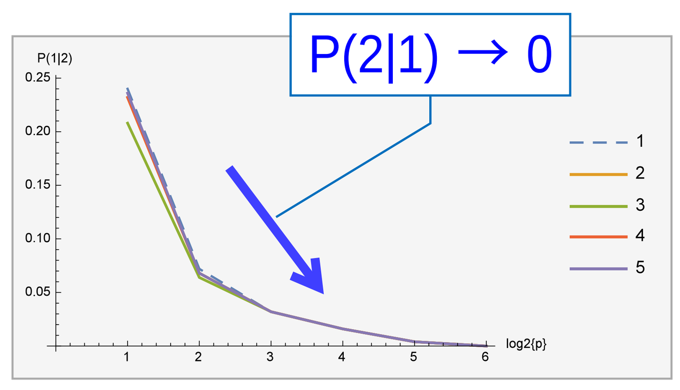


(a) $P(2|1)$ for the statistics 1, 2, 3, 4, and 5

Figure 1: The misclassification rates in simulations (a). The disturbance processes of $\pi_i, i = 1, 2$ are AR and MA process, respectively.



(b) $P(1|2)$ for the statistics 1, 2, 3, 4, and 5



(b) $P(2|1)$ for the statistics 1, 2, 3, 4, and 5

Figure 2: The misclassification rates in simulations (b). The disturbance processes of $\pi_i, i = 1, 2$ are AR and ARMA process, respectively.

2 Cluster analysis of financial data

- Data: 42 time series with 15 years in a company; Profit and Loss statement, Balance Sheet, and Cash Flow from the first and second section companies of Tokyo Stock Exchange.
- Disparity of the cluster analysis:

$$C(\boldsymbol{X}_1, \boldsymbol{X}_2) = (\boldsymbol{X}_1 - \boldsymbol{X}_2)'(\boldsymbol{X}_1 - \boldsymbol{X}_2).$$

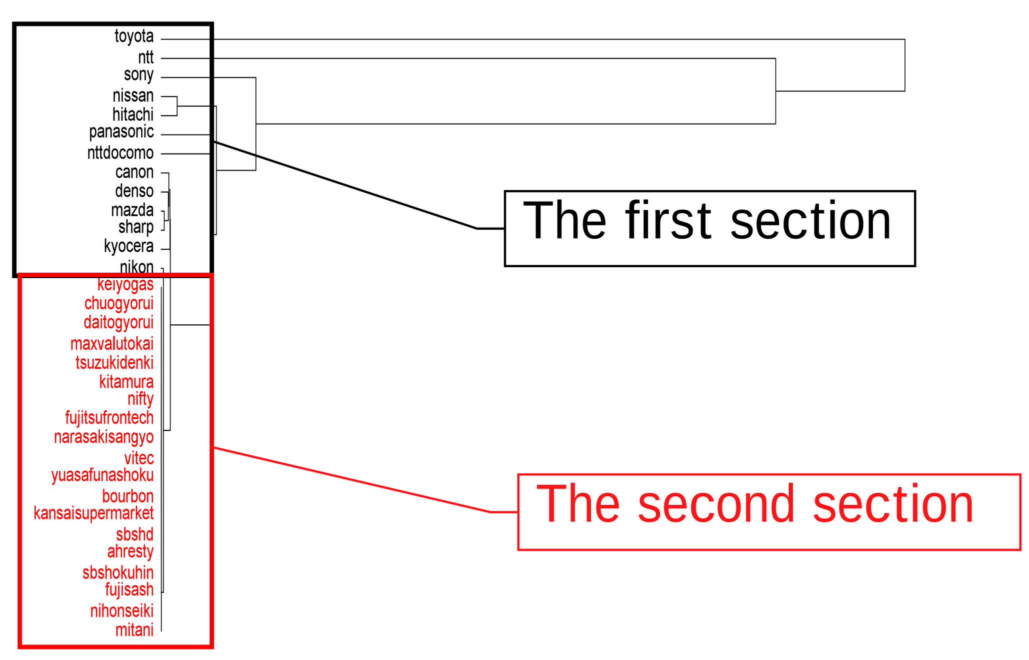


Figure 3: Cluster analysis of the first & second sections.