科研費シンポジウム

「多様な分野における統計科学の新展開」

日時：日時：2015 年 10 月 24 日（土）～10 月 26 日（月）

場所：富山県民会館　７０４号室

（TEL:０７６－４３２－３１１１）

http://www.bunka-toyama.jp/kenminkaikan/facilityguide/index.html

科学研究費・基盤研究（A)（課題番号：15H01678）

「大規模複雑データの理論と方法論の総合的研究」

研究代表者：青嶋 誠（筑波大学）

開催責任者：蛭川 潤一（新潟大学），星野 伸明（金沢大学）

Program

Saturday, 24, October

Reception 13:00-13:10

Opening 13:10-13:15 **Junichi HIRUKAWA** (Niigata University)

Afternoon Session I 13:15-15:15
Chair **Junichi HIRUKAWA** (Niigata University)
1. 13:15-13:55 **大石 惇喜 (Atsunobu Oishi)**, 白石 博 (Hiroshi Shiraishi)
慶應義塾大学理工学研究科 (Graduate School of Mathematical Sciences, Keio University)
慶應義塾大学理工学部 (Department of Mathematics, Keio University)
「最適配当境界の統計的推定」(Statistical Estimation for Optimal Dividend Barrier)
abstract
2. 13:55-14:35 **細野 剛二郎 (Tsuyojiro Hosono)**, 加藤 剛 (Takeshi Kato)
上智大学大学院理工学研究科数学領域・院 (Mathematics, Graduate School of Science and Technology, Sophia University)
上智大学理工学部情報理工学科 (Department of Information and Communication Sciences, Faculty of Science and Technology, Sophia University)
「重複ありのウェーブレット変換を利用した原油先物のリスクに関する考察」
(A risk analysis for crude oil futures by nondecimated discrete wavelet transform)
abstract
3. 14:35-15:15 **若林 光太 (Kota Wakabayashi)**, 加藤 剛 (Takeshi Kato)
上智大学大学院理工学研究科数学領域・院 (Mathematics, Graduate School of Science and Technology, Sophia University)
上智大学理工学部情報理工学科 (Department of Information and Communication Sciences, Faculty of Science and Technology, Sophia University)
「ウェーブレット分散を利用した原油先物のリスクに関する考察」
(A risk analysis for crude oil futures by wavelet variance)
abstract

Coffee Break 15:15-15:30

Afternoon Session II 15:30-17:30
Chair **Takayuki Shiohama** (Tokyo University of Science)
4. 15:30-16:10 **上原 悠槙 (Yuma Uehara)**
九州大学大学院数理学府 (Graduate School of Mathematics, Kyushu University)
「エルゴード的レヴィ駆動型確率微分方程式の段階的推定」
(Stepwise estimation of ergodic Levy driven stochastic differential equation)
abstract
5. 16:10-16:50 **布能 英一郎 (Eiichiro Funo)**
関東学院大学経済学部 (School of Economics, Kanto Gakuin University)

「Pooling incomplete samples を伴う多変量離散分布における Kullback 情報量の直和分解について」
(Decomposition of the Kullback information in discrete multivariate distributions by pooling incomplete samples)
abstract

6. 16:50-17:30 **新村 秀一 (SHUICHI SHINMURA)**
成蹊大学経済学部 (Faculty of Economics, Seikei University)
「判別関数の誤分類確率と判別係数の 95 %信頼区間」
(The 95% confidence intervals of error rates and discriminant coefficients)
abstract

Sunday, 25, October

Morning Session 9:10-11:50
Chair **Takeshi Kato** (Sophia University)
7. 9:10-9:50 **Simon Clinet**
The uiversity of Tokyo, grad. school of Mathematical sciences
Statistical inferences for ergodic point processes and applications to Limit Order Book
abstract
8. 9:50-10:30 **POTIRON Yoann**
The University of Chicago, Statistics Department
Parametric time series analysis and forecasting while taking proper account of the possible non-constancy of the underlying parameters of the model
abstract
9. 10:30-11:10 **Muneya Matsui**
Nanzan University
The extremogram and the cross-extremogram for a bivariate GARCH（1,1）process
abstract
10. 11:10-11:50 **金川 秀也 (Shuya Kanagawa)**
東京都市大学 共通教育部 (Department of Mathematics, Tokyo City University)
「株価日次収益率分析におけるジャンプ拡散過程モデルの同定とそのジャンプ時点推定への応用について」
(Estimation of Jump-Times of daily share prices of Nikkei 225 Stock Index using a jump diffusion model)
abstract

Lunch 11:50-13:15

Afternoon Session I 13:00-15:40
Chair **Yoshihiko MAESONO** (Kyushu University)
11. 13:15-13:55 **小部 敬純 (Takasumi KOBE)**, 種市 信裕 (Nobuhiro Taneichi), 関谷 祐里 (Yuri Sekiya)
鹿児島大学大学院 理工学研究科 (Graduate School of Science and Engineering KAGOSHIMA UNIVERSITY)
鹿児島大学・理工 (Department of Mathematics and Computer Science, Graduate School of Science and Engineering, Kagoshima University)
北海道教育大学・釧路 (Kushiro Campus, Hokkaido University of Education)
「3 次元分割表における 1 要因対 2 要因の独立性検定による改良変換統計量」
(Improved transformed statistics for the test of one factor independence from the other two in an 3 ways contingency table)
abstract
12. 13:55-14:35 **西山 貴弘 (Takahiro Nishiyama)**, 山田 雄紀 (Yuki Yamada), 兵頭 昌 (Masashi Hyodo)
専修大学 経営学部 (Department of Business Administration, Senshu University)

東京理科大学大学院 理学研究科 (Graduate School of Science, Tokyo University of Science)
大阪府立大学 工学部 (Department of Mathematical Sciences, Osaka Prefecture University)
「高次元枠組みにおける共分散構造に関する検定について」
(On a test for covariance structure in high-dimensional settings)
abstract
13. 14:35-15:15 **牛嶋 大 (Masaru Ushijima)**
（公財）がん研究会ゲノムセンター (Genome Center, Japanese Foundation for Cancer Research)
「Dirichlet Process を用いた遺伝子発現データのクラスタリング」
(Clustering of gene expression data using Dirichlet process model)

Coffee Break 15:15-15:30

Afternoon Session II 15:30-17:30
Chair **Hiroshi Shiraishi** (Keio University)
14. 15:30-16:10 阿部 俊弘 (Abe Toshihiro), 小方 浩明 (Ogata Hiroaki), **塩濱 敬之 (Takayuki Shiohama)**, 谷合 弘行 (Taniai Hiroyuki)
南山大学・理工 (Department of Mathematical Science, Nanzan University)
首都大学・都市教養 (Department of Business Administration, Tokyo Metropolitan University)
東京理科大学 ・工 (Department of Management Science, Tokyo University of Science)
早稲田大学・国際教養 (School of International Liberal Studies, Waseda University)
「時変自己相関を持つ円周上のマルコフ過程」
(Modeling Circular Markov Processes with Time Varying Autocorrelation)
abstract
15. 16:10-16:50 **島谷 健一郎 (Kenichiro Shimatani)**
統計数理研究所 (The Institute of Statistical Mathematics)
「非定常空間クラスター点過程におけるパラメータ推定：田中・尾形の Palm 尤度法の拡張」
(Statistical inference for nonstationary cluster point processes: an extension of Tanaka-Ogata's Palm likelihood method)
abstract
16. 16:50-17:30 **張 元宗 (Yuan-Tsung Chang)**
目白大学 (Mejiro University)
「順序制約がある２つの正規母平均の推定－分散共分散が既知の場合」
(Estimation of Two Ordered Normal Means with Known Covariance Matrix)
abstract
Conference Dinner 18:30-20:30
パレブラン高志会館　旬彩千歳
(PALAIS BLANK KOSHIKAIKAN Shunsai Chitose)
5,000 yen
(TEL: ０７６－４４１－２２５５)
`http://www.koshikaikan.com/information/2015/09/post-67.html`
map

Monday, 26, October

Morning Session 9:10-11:50
Chair **Nobuaki Hoshino** (Kanazawa University)
17. 9:10-9:50 **森山 卓 (Taku MORIYAMA)**, 前園 宜彦 (Yoshihiko MAESONO)
九州大学大学院数理学府 (Graduate School of Mathematics, Kyushu University)
九州大学大学院数理学研究院 (Faculty of Mathematics, Kyushu University)
「比のカーネル型推定量の漸近的性質について」
(Asymptotic properties of kernel estimators of ratios)
abstract
18. 9:50-10:30 **石井 晶 (Aki Ishii)**
筑波大学数理物質科学研究科 (Graduate School of Pure and Applied Sciences, University of Tsukuba)
First principal component and its applications to tests of means and covariance matrices for high-dimensional data
abstract
19. 10:30-11:10 **関谷 祐里 (Yuri Sekiya)**, 種市 信裕 (Nobuhiro Taneichi)
北海道教育大学・釧路 (Kushiro Campus, Hokkaido University of Education)
鹿児島大学・理工 (Department of Mathematics and Computer Science, Graduate School of Science and Engineering, Kagoshima University)
「漸近展開の不連続項を利用した離散バートレット型変換統計量の性質について」
(On the properties of discrete Bartlett-type transformed statistics using a discontinuous term of asymptotic expansion)
abstract
20. 11:10-11:50 **種市 信裕 (Nobuhiro Taneichi)**, 関谷 祐里 (Yuri Sekiya), 外山 淳 (Jun Toyama)
鹿児島大学・理工 (Department of Mathematics and Computer Science, Graduate School of Science and Engineering, Kagoshima University)
北海道教育大学・釧路 (Kushiro Campus, Hokkaido University of Education)
数学利用研究所 (The Institute for the Practical Application of Mathematics)
「二項反応における一般化線型モデルのリンク関数の拡張」
(On an extension of link functions for binomial generalized linear models)
abstract

Closing 11:50-11:55 **Makoto Aoshima** (University of Tsukuba)

最適配当境界の統計的推定
(Statistical Estimation for Optimal Dividend Barrier)

慶應義塾大学理工学研究科 大石惇喜 (Atsunobu Oishi)
慶應義塾大学理工学部 白石博 (Hiroshi Shiraishi)

1 Introduction

保険会社の破産リスクに関する理論 (risk theory, ruin theory) の応用として, 会社の余剰資本 (surplus) がある境界 (barrier) を上回ったときに, その部分を株主に返還する配当 (dividend) の問題がある [De Finetti, 1957]. 本研究では, この最適な境界 (optimal dividend barrier) の推定問題を考える.

配当が無い場合，時刻 $t(\geq 0)$ における保険会社のサープラス $U(t)$ が次のように表されているとする [Cramér, 1930].

$$U(t) = u + ct - S(t), \quad S(t) = \sum_{i=1}^{N(t)} X_i \tag{1}$$

ここで，u とは初期サープラスを表し，非負の定数とする（つまり，$u \geq 0$）．また，c とは単位時間当たりの保険料率を表し，正の定数とする．$S(t)$ とは，時刻 t における累積保険金額を表し，右式のような複合ポアソン過程に従うとする．さらに，$N(t)$ は時刻 t における保険金請求の累積頻度を表し，強度パラメータ $\lambda(>0)$ のポアソン過程に従うとし，$X_i(i \in \mathbb{N})$ は第 i 番目に請求された保険金額を表し，ある正値をとる確率変数に従うと仮定する．本論文では，特に，$\{X_i\}$ は i.i.d.(独立同一分布に従う) の（連続型）確率変数列とし，その確率密度関数 $f(x)$ は**未知**であると仮定する．

次に，上記のモデルに対し，配当の影響を考慮したものに修正する．具体的には，ある境界 $b(\geq u)$ を設定し，サープラスが b に到達した場合，次の保険金の請求が起こるまでの保険料収入を配当として契約者に還元するように修正する．このようなモデルでは，$D_b(t)$ を時刻 t における累積配当金額とすると，微小時間における $D_b(t)$ の変化率は次のように表すことができる．

$$dD_b(t) = \begin{cases} 0 & \text{if } U_b(t) < b \\ cdt & \text{if } U_b(t) = b \end{cases} \tag{2}$$

ここで，$U_b(t)$ とは配当境界 b を設定した場合の時刻 t におけるサープラスであり，次のように表すことができる.

Remark 1. $\{\Delta_i T, i \in \mathbb{N}\}$ を *i.i.d.* 確率変数列で母数 $\lambda > 0$ の指数分布に従っているとし，$T_i = \sum_{k=1}^{i} \Delta_k T,\ T_0 = 0$ とおく．このとき，各 $i = 1, 2, \ldots$ について，時刻 $t \in [T_{i-1}, T_i)$ におけるサープラス $U_b(t)$ は

$$U_b(t) = \begin{cases} U_b(T_{i-1}) + ct & \text{if } T_{i-1} \leq t < \min\{T_{i-1} + (b - U_b(T_{i-1}))/c, T_i\} \\ b & \text{if } \min\{T_{i-1} + (b - U_b(T_{i-1}))/c, T_i\} \leq t < T_i \end{cases}$$

である．ここで，$U_b(T_i-)$, $U_b(T_i)$ は次の通りである.

$$U_b(T_i-) = \min\left\{U_b(T_{i-1}) + c\Delta_i T, b\right\}, \quad U_b(T_i) = \begin{cases} u & \text{if } i = 0 \\ U_b(T_i-) - X_i & \text{if } i = 1, 2, \ldots \end{cases}$$

最適な配当境界問題として，次式で定義される “累積配当金現在価値の期待値”（$V(u, b)$）を最大にするような b を最適配当境界（optimal dividend barrier）とする.

$$V(u, b) = E\left[\int_0^{T_b} e^{-\delta t} dD_b(t)\right] \tag{3}$$

ここで，$T_b = \inf\{t | U_b(t) < 0\}$ とは破産時刻を表し，$\delta(>0)$ は割引率を表している.

この累積配当金現在価値 $A_b := \displaystyle\int_0^{T_b} e^{-\delta t} dD_b(t)$ は次のように表すことができる.

Remark 2. $\{X_i\}$ を（1）で定義した i.i.d. 確率変数列とし，$\{\Delta_i T\}$ を Remark 1 で定義した i.i.d. 確率変数列とする．また，$\{X_i\}$ と $\{\Delta_i T\}$ は互いに独立であるとする．このとき，累積配当金現在価値 A_b は

$$A_b = \frac{c}{\delta} \sum_{i=1}^{N(T_b)} e^{-\delta T_i} \left(e^{\delta \Delta_i \tau_b} - 1\right)$$

と表すことができる．ここで，$\Delta_i \tau_b$, T_b は次の通りである.

$$\Delta_i \tau_b = \max\left\{\Delta_i T - \frac{b - U_b(T_{i-1})}{c}, 0\right\}, \quad T_b = \inf\left\{t | U_b(t) < 0\right\}$$

今後，最適配当境界を $b^* \equiv b^*(u)$ と表す．即ち，ある定数 $u > 0$ を固定した上で，b^* は次式で定義される.

Definition 1.

$$b^* := \arg\max_{b \geq u} V(u, b) \tag{4}$$

2　推定問題

Remark 2 で用いた i.i.d. 確率変数列 $\{(X_i, \Delta_i T) : i \in \mathbb{N}\}$ の部分列として $\mathbb{Y}_n := \{(X_i, \Delta_i T) : i = 1, \ldots, n,\ n \in \mathbb{N}\}$ が観測されているとする. このデータ集合 $\mathbb{Y}_n$ から，我々は次のように，別の観測列を生成することができる.

$$\mathbb{Y}_N^{(j)} := \{(X_i^{(j)}, \Delta_i T^{(j)}) : i = 1, \ldots, N,\ (X_i^{(j)}, \Delta_i T^{(j)}) \in \mathbb{Y}_n\}$$

上記で生成した観測列 $\mathbb{Y}_N^{(j)}$ を同じように M 個生成することを考える.

$$\mathbb{Y}_N^{(1)}, \ldots, \mathbb{Y}_N^{(M)}$$

このようにして生成した観測列 $\mathbb{Y}_N^{(j)} = \{(X_i^{(j)}, \Delta_i T^{(j)})\}$ を 1 つ固定し，ある固定した b に対応したサープラス過程の観測列 $\{U_b^{(j)}(t)\}$ を次のように定義する:

$$U_b^{(j)}(t) = \begin{cases} \min\left\{U_{b,0}^{(j)} + ct, b\right\} & \text{if } 0 \le t < T_1^{(j)} \\ \min\left[U_{b,1}^{(j)} + c\left\{t - T_1^{(j)}\right\}, b\right] & \text{if } T_1^{(j)} \le t < T_2^{(j)} \\ \vdots & \vdots \\ \min\left[U_{b,i-1}^{(j)} + c\left\{t - T_{i-1}^{(j)}\right\}, b\right] & \text{if } T_{i-1}^{(j)} \le t < T_i^{(j)} \\ \vdots & \vdots \\ \min\left[U_{b,N-1}^{(j)} + c\left\{t - T_{N-1}^{(j)}\right\}, b\right] & \text{if } T_{N-1}^{(j)} \le t \le T_N^{(j)} \end{cases}$$

$$\text{ここで, } T_i^{(j)} = \sum_{k=1}^{i} \Delta_k T^{(j)}, \quad U_{b,i}^{(j)} = \begin{cases} u & \text{if } i = 0 \\ U_{b,i}^{(j)-} - X_i^{(j)} & \text{if } i = 1, 2, \ldots, N \end{cases}, \quad U_{b,i}^{(j)-} = \lim_{t \uparrow T_i^{(j)}} U_b^{(j)}(t)$$

さらに (擬似的な) 破産時刻 $T_b^{(j)}$, 区間 $[T_{i-1}^{(j)}, T_i^{(j)})$ で $U_b^{(j)}(t) = b$ となっている期間 $\Delta_i \tau_b^{(j)}$, 区間 $[0, T_b^{(j)}]$ で発生する保険金支払回数 $N(T_b^{(j)})$ を以下で定義する.

$$T_b^{(j)} = T_N^{(j)} \wedge \inf\{t | U_b^{(j)}(t) < 0\}, \quad \Delta_i \tau_b^{(j)} = \min\left\{\Delta_i T^{(j)} - \frac{b - U_{b,i-1}^{(j)}}{c}, 0\right\},$$

$$N(T_b^{(j)}) = \begin{cases} N & \text{if } T_b^{(j)} = T_N^{(j)} \\ \min\{\ell | U_b^{(j)}(T_\ell^{(j)}) < 0\} & \text{if } T_b^{(j)} = \inf\{t | U_b^{(j)}(t) < 0\} \end{cases}$$

このような設定の上で，$V(u, b)$ の推定量を次のように定義する.

Definition 2.

$$\widehat{V}_n(u, b) := \frac{1}{M} \sum_{j=1}^{M} \hat{A}_b^{(j)} := \frac{1}{M} \sum_{j=1}^{M} \frac{c}{\delta} \sum_{i=1}^{N(T_b^{(j)})} e^{-\delta T_i^{(j)}} (e^{\delta \Delta_i \tau_b^{(j)}} - 1)$$

このとき, Definition 1 で表した b^* の推定量を次のとおり定義する.

Definition 3.

$$\hat{b}_n^* = \underset{b \ge u}{\operatorname{argmax}} \widehat{V}_n(u, b)$$

このとき, $\hat{b}_n^*$ に関する一致性が成り立つことを報告した.

Theorem 1. サンプルサイズ n が大きくなると同時に，リサンプリングサイズ $N \equiv N(n)$ および（再）標本数 $M \equiv M(n)$ も同時に大きくなるとする．つまり，

$$N(n) \to \infty, \quad M(n) \to \infty \qquad \text{as} \quad n \to \infty$$

が成り立つとする．このとき次が成り立つ.

$$\hat{b}_n^* \xrightarrow{p} b^*$$

参考文献

[De Finetti, 1957] De Finetti, B. (1957). Su un fimpostazione alternativa della teoria collettiva del rischio. In *Transactions of the XVth international congress of Actuaries*, volume 2, pages 433–443.

[Cramér, 1930] Cramér, Harald. (1930) On the mathematical theory of risk.

[Van der Vaart, 2000] Van der Vaart, A. W. (2000) Asymptotic statistics (Vol. 3). Cambridge university press.

重複ありのウェーブレット変換を利用した原油先物のリスクに関する考察

上智大学大学院理工学研究科数学領域・院　細野 剛二郎
上智大学理工学部情報理工学科　加藤　剛

1　ウェーブレット変換

1.1　連続ウェーブレット変換

ウェーブレット変換は, ある条件を満たすウェーブレットとよばれる関数を用いて, 様々な関数を成分ごとに分解する変換で, 関数の特徴を調べることができる. 特に，関数の局所的な挙動を調べることに威力を発揮する. 伝統的に用いられてきたフーリエ変換と比較すると，時系列データを対象にした場合，フーリエ変換が定常時系列の周波数情報を抽出する道具であるのに対し，ウェーブレット変換は，時系列の変化点検出や時変周波数の推定に用いられる.

具体的なウェーブレットの例としては, マザーウェーブレットを

$$\psi(x) = \begin{cases} 1 & x \in [0, \frac{1}{2}), \\ -1 & x \in [\frac{1}{2}, 1), \\ 0 & \text{otherwise}, \end{cases}$$

としたものから生成されるハールウェーブレットがある. ハールウェーブレットは最も簡潔なウェーブレットであるが，その簡潔さによる計算の容易さと，簡素であっても意味のある結果を導く事例が多数報告されていることから，実際の解析でよく用いられる. 本研究においても，ハールウェーブレットを利用する.

1.2　離散ウェーブレット変換

離散データ $y = (y_1, \ldots, y_n)$ に対するハールウェーブレット変換を考える. このとき, 離散ウェーブレット変換の性質の都合上, n は 2 のべき乗とし, $n = 2^J$ とする. そして, 2 種類の係数,

$$\begin{cases} d_{j,k} = (c_{j+1,2k} - c_{j+1,2k-1})/\sqrt{2} \\ c_{j,k} = (c_{j+1,2k} + c_{j+1,2k-1})/\sqrt{2} \end{cases}$$

特に, $j = J$ のときは,

$$\begin{cases} d_{J,k} = (y_{2k} - y_{2k-1})/\sqrt{2} \\ c_{J,k} = (y_{2k} + y_{2k-1})/\sqrt{2} \end{cases}$$

と定義する. ただし, j と k はそれぞれ, $0 \le j \le J$, $1 \le k \le 2^j$ をみたす整数である. $\{c_{j,k}\}$ と $\{d_{j,k}\}$ を（ハールウェーブレットによる）ウェーブレット係数と呼ぶ. 連続の場合と同様に，各 j, k におけるウェーブレット係数の値から，時系列の変化点検出や時変周波数の推定を行うことができる.

1.3　重複ありの離散ウェーブレット変換

離散ウェーブレット変換の問題点としては, j が小さいとき, つまり周期が大きいときに係数の数が少なくなるため, データの変化を検出しづらいことである. この欠点を補うものが重複ありの離散ウェーブレット変換であり, 2 種類の係数をそれぞれ,

$$\begin{cases} d_{j,k} = (c_{j+1,k} - c_{j+1,k-1})/\sqrt{2} \\ c_{j,k} = (c_{j+1,k} + c_{j+1,k-1})/\sqrt{2} \end{cases}$$

特に, $j = J$ のときは,

$$\begin{cases} d_{J,k} = (y_k - y_{k-1})/\sqrt{2} \\ c_{J,k} = (y_k + y_{k-1})/\sqrt{2} \end{cases}$$

で定義する. ただし, j と k はそれぞれ, $0 \le j \le J$, $1 \le k \le n$ をみたす整数である.

2　原油先物価格の分析

金融商品の価格の時系列データを重複ありの離散ウェーブレット変換をすることにより, 価格が大きく変動する際の前兆現象をつかむことができるのではないかと考えた. 例えば原油先物を対象にした場合，もしも大きな下落の前兆現象を何らかの数値指標の形でとらえることができれば，大口需要家や投資家は，先物取引を手じまいしたり，資産内容の組み替えを行ったりして，損失を回避または減少させることが可能になる.

本研究では，東京商品取引所が扱う原油先物の約定値段のデータを 15 分次データに換算し，重複ありの離散ウェーブレット変換を用いて分析した.

2.1　2014 年の原油先物価格の暴落

図 1 は 2014 年 8 月 8 日から 10 月 31 日までの東京商品取引所の原油先物価格（約定値段）の動きを表すもので, データの個数は $4096 (= 2^{12})$ 個である. 横軸 2600 付近の縦線は 10 月 1 日 20 時 45 分であり, その後から値を大きく下げ始めている.

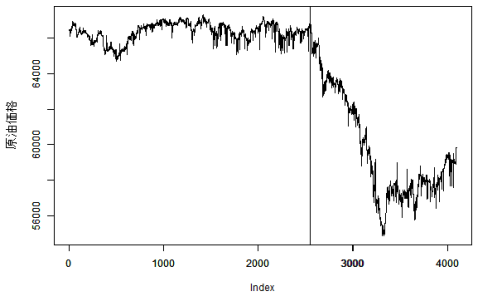


図 1　2014 年の原油先物価格

このデータに対して重複ありの離散ウェーブレット変換を行うと，レベル 0 から 11 までのウェーブレット係数が得られる．その係数は，価格の下落を反映して，小さいレベルから中程度のレベルで下落後に大きく振れていることが見て取れる．また，小さいレベルと比較して，大きいレベルの係数は絶対値が相対的に小さくなるという, 離散ウェーブレット変換の性質が現れている．

一番高いレベル 11 の係数の自乗を, それぞれの点から前に, その点も含めて 40 個の点の平均をとって平滑化したものが図 2 である．1500 あたりから 20000（図の中の横線）を超えていることがわかる．

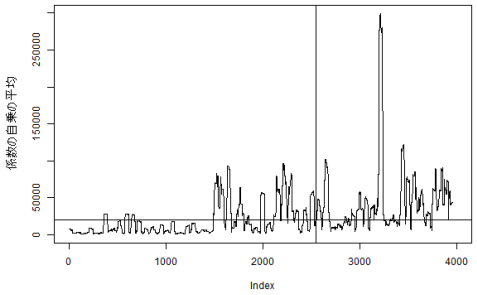


図 2　レベル 11 の係数の 2 乗の平均

2.2　リーマンショック前の価格上昇

図 1 とは反対の高騰する局面についても検討した．

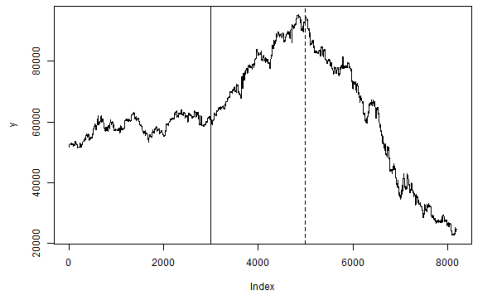


図 3　2007 年から 2008 年の原油先物価格

図 3 は 2007 年 9 月 18 日 14 時 45 分から 2008 年 12 月 30 日までの東京商品取引所の原油先物価格で, データの個数は 8192($= 2^{13}$) 個である．

実線は 2008 年 3 月 31 日 (横軸 3000) で，その後から価格が大きく上がっている．また, 点線は 2008 年 7 月 14 日 (横軸 5000) で，その後から大きく値を下げている．

このデータのウェーブレット係数を計算すると，値上がり局面になった後（図 4 の実線と点線の間）で中程度以上のレベルで係数が下に振れていることがわかった．そして，値下がりに転じた後（図 4 の点線以降）では，同じレベルで係数が上に振れていることも判明した．

図 2 と同じ方法でウェーブレット係数を平滑化したものが図 4 である．

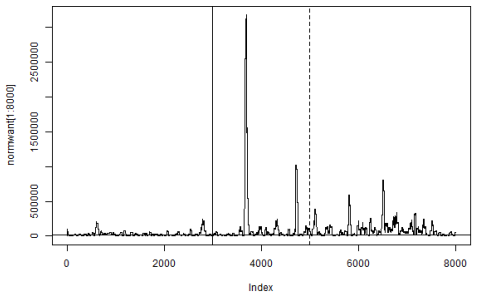


図 4　レベル 12 のウェーブレット係数の 2 乗の平均

値上がり後からは一貫して大きな値をとっている．また，値上がり前の 横軸 2800 付近で，それ以前より大きな値をとっている．

2.3　まとめ

図 2 と図 4 で示したように，価格が大きく動く前に, 短い周期の係数が大きく振れることがわかった．別の時刻における急騰や暴落の状況で更に検討を重ね, もしも短い周期のウェーブレット係数が大きく振れる特徴が現れるならば，それを前兆現象の 1 つとして利用できる可能性が出てくる．その場合，何が理由となって原油先物価格が短期間で上下するように変化するのかを考える必要もある．

参考文献

[1] G.P. Nason, *Wavelet Methods in Statistics with R*. Springer. (2008).

[2] G.P. Nason, R. Sachs, and G. Kroisandt, Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *J. R. Stat. Soc. Ser. B*, 62, 271-292 (2000).

[3] データ提供　東京商品取引所

ウェーブレット分散を利用した原油先物のリスクに関する考察

上智大学大学院理工学研究科数学領域・院　若林光太
上智大学理工学部情報理工学科　加藤　剛

1　研究の背景と目的

石油探査目的の人工地震波の解析の方法として1980 年代初めに考案されたウェーブレット解析は，工学分野を中心に急速な発展を遂げ，1990 年代終わり頃には数学的な理論的枠組みもほぼ完成した．その後，R を含む各種データ解析用ソフトウェアへの実装も進み，現在では，工学分野を中心に，フーリエ解析と同様にごく普通の道具になっている．

本研究は，ウェーブレット解析の中でもウェーブレット分散と呼ばれる概念を活用し，東京商品取引所 (TOCOM) における原油先物価格について，大きな下落や上昇の前兆現象を捉えることを目的としたものである．最終的には，下落や上昇の注意喚起や警報に使用できる数値指標を作ることを目標にしている．

先物取引では，商品相場の変動によって利益を得ることも損失を被ることもある．さらに，損失を生じたときに取引を継続する場合，その損失額によっては追証拠金という新たな負担が生じる．特に価格の暴落の前兆現象をつかむことができれば，大口需要家や投資家は，売りと買いを入れ替えたり，資金を他の金融商品に移動したり，手持ちの別の金融資産を売却して追証拠金の手当をしたりするなどして，損失が発生する状況にあらかじめ対処できる．また，適切なリスク指標を提供することは，市場の活性化にも繋がる．TOCOM 取り扱いの先物を対象にするとき，相場が急激に変動した際の措置が 2009 年 5 月に値幅制限制度からサーキットブレーカー制度へ変更されたこともあって，理論的裏付けをもったリスク評価に関する研究は少ない．

原油価格は，その標準指標である WTI (West Texas Intermediate) について，2014 年の半ばの 100 ドルから下落を続け，今年の 9 月末時点では 45 ドル前後にまで落ち込んでいる．2014 年 11 月の OPEC 総会で減産合意に失敗したときは，TOCOM で原油先物価格が暴落し，サーキットブレーカーが発動された．価格の下落局面が最近頻発してデータが集めやすいという理由により，本研究では原油先物を対象にした．

2　ウェーブレット分散

ウェーブレット分散とは，時系列の分散を分解することで時系列に対する分散分析を可能にするものである．時系列解析において最も広く用いられている分散分析の手法は，スペクトル解析である．フーリエ変換にもとづくスペクトル密度も，これに含まれる．ウェーブレット分散による分散分析は，多くの点においてスペクトル密度による分散分析と類似している．ところが，実際に使用する者の立場から見ると両者には重要な違いがある．

スペクトル密度は，連続的なフーリエ周波数の分散の分解である．分解によって得られる各々の成分は，特定の周波数を持つ正弦曲線に時系列がどの程度似ているかを反映するものである．その一方で，ウェーブレット分散は，スケールの離散集合の分解を与える．ここで，大まかな言い方をすれば，スケールとは時系列の平均値を取る時間の区間のことである．ウェーブレット分散による分解の各成分の長所は，ある特定のスケールにおける隣接した平均値の間にどれくらいの相違があるかを測れることにある．スケールという概念は，周期（周波数の逆数）の概念とは別のものである．スケールも周期も同じ単位で測れるものであるが，後者は平均をとるという考えを伴わない．

3　商品先物取引

本節における説明は，東京商品取引所のホームページ [5] に負う．商品先物取引とは，将来の一定期日に一定数の商品を契約時に取り決めた価格で買うまたは売ることを約束する取引のことで，金融派生商品の 1 つである．将来における売買の値段を予め決めておくため，商品価格相場の上下動次第では利益を得ることもあれば損失を被ることもある．希望を表明すれば商品を授受することは可能であるが，契約時に決定された期日までに当初の取引と反対の取引を行うことで，差金決済だけで取引を終了することもできる．

TOCOM における原油に関する種々の情報は，表 1 の形にまとめられる．

表 1：東京商品取引所の原油先物取引に関する情報

限月	新甫発会日の属する月の翌月から起算した6月以内の各月（6限月制）
立会時間	日中：9:00 〜 15:15 夜間：16:30 〜 翌日4:00
取引最終日	当限月の最終営業日（日中立会まで）
最終決済日	当限月の翌月第一営業日

4　数値実験

4.1　原油先物の価格変動

TOCOM における原油先物取引について，2014 年 11 月 1 日 16 時 30 分 00 秒から 11 月 30 日 15 時 15 分 00 秒までの約定値段の推移をグラフにしたものが図 1 である．ただし，上記の時刻表示は TOCOM の定める計算区域にもとづく時刻表記であり，以降も同様とする．

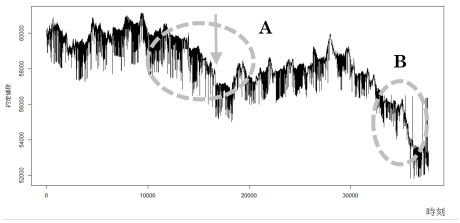


図 1: 2014 年 11 月 1 日 16 時 30 分 00 秒から 11 月 30 日 15 時 15 分 00 秒までの原油の約定値段

A と記した点線丸囲み部分の中央よりやや右の箇所で不連続的な値段の下落が起きている．下落直前の値動きを観察すると，比較的密な動きから隙間が空いた状態に移っていることがわかる．また，B と記した点線丸囲みの箇所では，不連続的とは言えないまでも，かなり急激な下落が生じている．下落に至る直前の動きを見ると，A の箇所ほど顕著ではないが，やはり比較的密な動きから隙間が空いた状態への遷移が見られる．ウェーブレット分散を活用して，これらの視覚的な値動きを数値の情報，あるいは，その数値情報を図にしたものとして抽出することを試みる．ウェーブレット分散の値について A と B に共通するものを得ることができれば，それが下落の前兆を示す情報として使える可能性がある．

4.2　結果

図 1 の A の部分について，矢印の示す不連続的な下落の 1 つ前のデータから 4096 個のデータをさかのぼって抽出したものが図 2 である．図 2 にあるように，下落の直前の区間を区間 1 とし，さかのぼって区間 2，区間 3，区間 4 をとる．そして，それぞれの区間におけるウェーブレット分散を計算した．

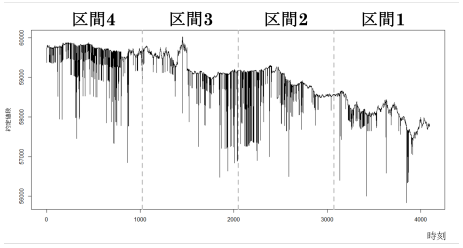


図 2: A の部分の拡大およびウェーブレット分散を計算する区間

結果を図 3 に示す．下落直前の特徴は，水準 $j = 3, 4$ におけるウェーブレット分散の著しい減少である．$\tau_j = 2^{j-1}$ のスケール換算をすると，$\tau_3 = 4$, $\tau_4 = 8$ における減少である．そして，図 3 ほどに顕著ではないが，図 1 における B の下落箇所においても同様の結果が得られた．水準 $j = 3, 4$ におけるウェーブレット分散の減衰を見ることで，原油価格の暴落のサインを得られる可能性があることがわかった．

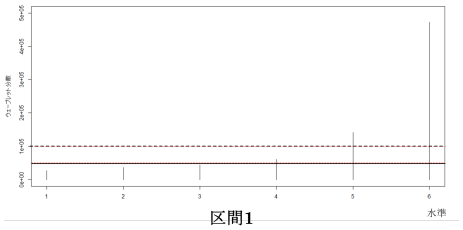


図 3: 区間 1 のウェーブレット分散

参考文献

[1] P.J. Brockwell and R. A. Davis, ”Time Series: Theory and Methods”, *Springer*, (1991).

[2] G.P. Nason, R. Sachs, and G. Kroisandt, Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *J. R. Stat. Soc. Ser. B*, 62, 271-292 (2000).

[3] Nason, ”Wavelet Methods in Statistics with R”, *Springer*, (2008).

[4] T. S. Rao, S. S. Rao, and C.R.Rao ”Handbook of Statistics, Volume 30: Time Series Analysis: Methods and Applications”, *ELSEVIER*, (2012).

[5] 東京商品取引所ホームページ
http://www.tocom.or.jp/jp/

Stepwise estimation of ergodic Lévy driven stochastic differential equation (エルゴード的Lévy駆動型確率微分方程式における段階的推定)

九州大学 上原 悠槙
九州大学 増田弘毅

現在, 観測機器の発達などにより金融データや電気信号といった時間発展現象の高頻度観測が可能となっているが, それらの観測において観測ノイズをどう設定するかはモデリングの観点から重要である. 時間発展現象の連続時間モデルとして, 正規性を有する連続なノイズ過程である Wiener 過程を用いた拡散過程が用いられてきた. しかし実データにおいては, ノイズが正規性を有さない場合が多く, またスパイクノイズといった不連続なノイズが確認されており, これらの表現として連続な要素と不連続な要素の両方を持つ Lévy 過程が近年用いられている. だが拡散過程とは異なり, その不連続性などから Lévy 駆動型確率微分方程式の統計的推測に関する先行研究は各論が主となっている. また drift 係数や scale 係数と駆動する Lévy ノイズの情報の同時推定方式はほぼ未確立である.

本シンポジウムにおいては, 広範の Lévy 駆動型確率微分方程式に適用可能な推定方式についての研究報告を行った. 以下に具体的内容を述べる. 我々は, エルゴード性を有する 1 次元確率微分方程式,

$$dX_t = a(X_t, \alpha, \gamma)dt + c(X_{t-}, \gamma)dJ_t, \quad X_0 = x_0,$$

を想定した. ここで

- $\theta := (\alpha, \gamma) \in \Theta \subset \mathbb{R}^p$ は有限次元未知パラメータ.
- drift 係数 $a : \mathbb{R} \times \Theta \to \mathbb{R}$ と scale 係数 $c : \mathbb{R} \times \Theta \to \mathbb{R}$ の関数形は既知 (すなわち係数に対する誤特定は考慮しない).
- 駆動ノイズ J_t は純粋ジャンプ型 Lévy 過程であり, 任意の $q > 0$ について $E[|J_1|^q] < \infty$ を満たす.

である. このような条件を満たすような J_t は, normal inverse Gaussian 過程や tempered stable 過程といった応用上極めて重要な Lévy 過程を包含していることを注意しておきたい. 上記の連続時間モデルからの高頻度観測 $(X_{t_0}, X_{t_1}, \ldots, X_{t_n})$ を想定する, ここで観測時間は $t_j = t_j^n = jh_n$ と表され, 正数 h_n は $nh_n \to \infty$, $nh_n^2 \to 0$ を満たすものとする. この観測 $(X_{t_0}, X_{t_1}, \ldots, X_{t_n})$ に基づき, 我々は以下の段階的推定法を考案した:

1. γ のドリフトフリー推定. $\hat{\gamma}_n$ を以下のドリフトフリーガウス型疑似尤度に基づいた, ランダム方程式の解とする.
$$G_{1,n}(\gamma) := \frac{1}{nh_n} \sum_{j=1}^{n} \frac{\partial_\gamma c_{j-1}(\gamma)}{c_{j-1}(\gamma)^3} \left\{ (\Delta_j X)^2 - h_n c_{j-1}^2(\gamma) \right\} = 0,$$
ここで $\Delta_j X = X_{t_j} - X_{t_{j-1}}$ であり, 関数 f について $f_j(\theta) = f(X_{t_j}, \theta)$ と表記している.
2. $\hat{\gamma}_n$ プラグインによる α の推定. $\hat{\alpha}_n$ を以下のランダム方程式の解とする.
$$G_{2,n}(\alpha, \hat{\gamma}_n) := \frac{1}{nh_n} \sum_{j=1}^{n} \frac{\partial_\alpha a_{j-1}(\alpha, \hat{\gamma}_n)}{c_{j-1}^2(\hat{\gamma}_n)} \left\{ \Delta_j X - h_n a_{j-1}(\alpha, \hat{\gamma}_n) \right\} = 0.$$

このようにして定義された段階的推定量 $\hat{\theta}_n$ について, その一致性, また漸近正規性の導出を行った:

$$\hat{\theta}_n \xrightarrow{P_0} \theta_0,$$
$$\sqrt{nh}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} N(0, \mathcal{I}^{-1}\Sigma(\mathcal{I}^{-1})^T).$$

また発表においては観測データのみによって漸近分散の不偏推定量を構成することにより, 標準漸近正規性を導き, パラメータの信頼区間の構成ができることも紹介した. 更に Masuda and Uehara [1] で用いられた駆動ノイズに対応する Lévy 測度の関数型推定の手法を適用することで, 本モデルにおいても Lévy 測度の関数型推定を同時に行えることも同様に述べた. Lévy 測度の推定は, ノイズの変動を捉える上で極めて重要であり, 時間発展現象のモデリングにおいて重要な役割を果たすことが期待される. 以上に述べた段階型推定法の利点として, (α, γ) の推定を分けることによる計算負荷の軽減, ノイズの誤特定に対する頑健性が挙げられ, 広範の Lévy 駆動型確率微分方程式の統計的推測手法が確立された.

また本シンポジウムにおいては, 他の講演者の方々との議論, 発表を通して多数の大規模データに対する最新のアプローチに触れることができた. 今後はそれらに基づき, より精密な方法論の構築だけでなく, 実社会に内在する大規模データの解析へのさらなる応用を模索していきたいと考えている.

References

[1] Masuda, H. and Uehara, Y. (2015). Two-step estimation of ergodic Lévy driven SDE, arXiv:1505.01922.

Pooling incomplete samples を伴う多変量離散分布における Kullback 情報量の直和分解について

関東学院大学経済学部　布能 英一郎

要旨 離散データ解析において、カテゴリーの減少が各セルの生起確率が比例配分で生じる場合に、統計的推測を行うのが Pooling incomplete samples である。特に Asano(1965) は、多項分布において pooling incomplete samples の下で、数々の数理的な成果を出した。

他方、情報理論の立場から離散データ解析を考察すると、多くの場合に Kullback 情報量の直和分解が成立する。特に、離散分布の 2 標本問題において、Between information と Within information の和が Total information に等しくなることが多くみられる。このことは、離散解析においても、分散分析と似た解析ができ、有益であると言える。

最近、Funo(2015) によって、多項分布における pooling incomplete samples の下での 2 標本問題において、Between information と Within information の和が Total information に等しいことが示された。

では、他の多変量離散分布の 2 標本問題に関して、Between information と Within information の和が Total information に等しくなるか？ 本研究は、この問題を、負の多項分布、Poisson 分布の積の場合に調べたものである。

Kullback 情報量の直和分解 多項分布の 2 標本問題 $(X_1^{[i]},\cdots,X_k^{[i]}) \sim Multi-nomial(N^{[i]}: p_1^{[i]},\cdots,p_k^{[i]})$, $i=1,2,\quad H_1: p_j^{[1]} \neq p_j^{[2]}$, $H_2: p_j^{[1]} = p_j^{[2]} = p_j$, $(j=1,\cdots,k)$ を考える。このとき、Kullback 情報量 $\hat{I}(p^*:p) = \sum_{i=1}^{2} N_1^{[i]} \sum_{j=1}^{k} x_j^{[i]} \log(x_j^{[i]}/N_1^{[i]} p_j)$ は、between $\hat{I}(\hat{p}:p) = \sum_{j=1}^{k}(x_j^{[1]}+x_j^{[2]}) \log((x_j^{[1]}+x_j^{[2]})/(N^{[1]}+N^{[2]})p_j)$ と within $\hat{I}(p^*:\hat{p}) = \sum_{i=1}^{2}\sum_{j=1}^{k} x_j^{[i]} \log((N^{[1]}+N^{[2]})x_j^{[i]}/N^{[i]}(x_j^{[1]}+x_j^{[2]}))$ に直和分解できる。(Kullback, 1968)

Pooling incomplete samples の場合 離散データ解析において、オリジナルな観測と、カテゴリー減少が生じて、その各セルの生起確率がオリジナルな観測のセル確率を比例配分とした観測との同時分布にもとづいて統計的推測を行うのが Pooling incomplete samples である。この場合の 2 標本問題において、Kullback 情報量の直和分解が成り立つこともあれば、成り立たないこともある。

定理 1. $\boldsymbol{X}^{[1]}, \boldsymbol{X}^{[2]}, \boldsymbol{Y}^{[1]}, \boldsymbol{Y}^{[2]}$ は互いに独立な多項分布 $\boldsymbol{X}^{[i]} = (X_1^{[i]},\cdots,X_m^{[i]},\cdots,X_k^{[i]})$, $\boldsymbol{Y}^{[i]} = (Y_1^{[i]},\cdots,Y_m^{[i]})$, $i=1,2$ で、$\boldsymbol{X}^{[i]} \sim Multinomial(N_1^{[i]}; p_1^{[i]},\cdots,p_m^{[i]},\cdots,p_k^{[i]})$, $\boldsymbol{Y}^{[i]} \sim Multinomial(N_2^{[i]}; p_1^{[i]}/\sum_{l=1}^{m} p_l^{[i]},\cdots,p_m^{[i]}/\sum_{l=1}^{m} p_l^{[i]})$ $i=1,2,\quad H_1: p_j^{[1]} \neq p_j^{[2]}$, $H_2: p_j^{[1]} = p_j^{[2]}$ とする。このとき、Kullback 情報量の直和分解が成り立つ。

負の多項分布の場合 さて、負の多項分布の場合、通常の 2 標本問題では Kullback 情報量の直和分解が成り立つものの、Pooling incomplete samples の場合、すなわち、定理 1. にて、分布の仮定を $\boldsymbol{X}^{[i]} \sim Negative\ Multinomial(r_1^{[i]}; p_1^{[i]},\cdots,p_m^{[i]},\cdots,p_k^{[i]})$,

$\boldsymbol{Y}^{[i]} \sim$ *Negative Multinomial*$(r_2^{[i]}; p_1^{[i]}/\sum_{l=0}^m p_l^{[i]}, \cdots, p_m^{[i]}/\sum_{l=0}^m p_l^{[i]})$ $i = 1, 2,$ とした場合、Kullback 情報量の直和分解が成り立たない。 他方、次の定理が得られた。

定理 2. $\boldsymbol{X}^{[1]}, \boldsymbol{X}^{[2]}, \boldsymbol{Y}^{[1]}, \boldsymbol{Y}^{[2]}$ は互いに独立で $\boldsymbol{X}^{[i]} \sim$ *Negative Multinomial*$(r_1^{[i]}; p_1^{[i]}, \cdots, p_m^{[i]}, \cdots, p_k^{[i]})$, $\boldsymbol{Y}^{[i]} \sim$ *Multinomial*$(N_2^{[i]}; p_1^{[i]}/\sum_{l=1}^m p_l^{[i]}, \cdots, p_m^{[i]}/\sum_{l=1}^m p_l^{[i]})$, $i = 1, 2,$ ならば、Kullback 情報量の直和分解が成り立つ。

Poisson の場合 $m < k$ とする。$X_j^{[i]}$ $(i = 1, 2,\ j = 1, 2, \cdots, k)$ および $Y_j^{[i]}$ $(i = 1, 2,\ j = 1, 2, \cdots, m)$ はすべて互いに独立で $X_j^{[i]} \sim$ *Poisson*$(\lambda_j^{[i]})$, $\quad i = 1, \cdots, k,$ $Y_j^{[i]} \sim$ *Poisson*$(\lambda_j^{[i]}(\sum_{l=1}^k \lambda_l^{[i]})/(\sum_{l=1}^m \lambda_l^{[i]})$, $\quad j = 1, \cdots, m,$ という確率モデルが、Poisson 分布における自然な pooling incomplete sample である。ところが、この場合、Total $\neq$ Between + Within である。他方、Poisson 分布の場合にも、次の定理が得られた。

定理 3. 各 $i = 1, 2$ に対して $X_j^{[i]} \sim$ *Poisson*$(\lambda_j^{[i]})$, $j = 1, \cdots, k,$ $\boldsymbol{Y}^{[i]} = (Y_1^{[i]}, \cdots, Y_m^{[i]}) \sim (N_2^{[i]}, \lambda_1^{[i]}/\sum_{l=1}^m \lambda_l^{[i]}, \cdots, \lambda_m^{[i]}/\sum_{l=1}^m \lambda_l^{[i]})$ であって $X_1^{[i]}, \cdots, X_k^{[i]}, \boldsymbol{Y}^{[i]}, i = 1, 2$ はすべて独立。このとき、Kullback 情報量の直和分解が成り立つ。

直和分解が成り立つ背景の考察 負の多項分布は 負の二項分布 × 多項分布 に分解される。実際、

$$\frac{(x_1 + \cdots + x_k + r - 1)!}{x_1! \cdots x_k! (r-1)!} p_0^r p_1^{x_1} \cdots p_k^{x_k} = \frac{(x_1 + \cdots + x_k + r - 1)!}{(x_1 + \cdots + x_k)!(r-1)!} p_0^r (1 - p_0)^{x_1 + \cdots + x_k}$$
$$\times \frac{(x_1 + \cdots + x_k)!}{x_1! \cdots x_m! \cdots x_k!} \left(\frac{p_1}{1 - p_0}\right)^{x_1} \cdots \left(\frac{p_m}{1 - p_0}\right)^{x_m} \cdots \left(\frac{p_k}{1 - p_0}\right)^{x_k}$$

そして、Pooling incomplete samples が「多項分布」の部分で行われていれば Total = Between + Within が成り立つ。しかし、他のところで Pooling incomplete samples が行われた場合は、等号が成立するとは限らない。

Poisson 分布の積の場合も、Poisson 分布の積 = 1 次元 Poisson × 多項分布 が成り立つ。実際、実際、パラメーター変換 $s = \sum_{i=1}^k \lambda_i,$ $u_i = \lambda_i/(\sum_{j=1}^k \lambda_j)$ により

$$\prod_{i=1}^k \frac{\lambda_i^{x_i} \exp(-\lambda_i)}{x_i!} = \cdots = \frac{s^{\Sigma_{i=1}^k x_i} \exp(-s)}{(\sum_{i=1}^k x_i)!} \frac{(\sum_{i=1}^k x_i)!}{x_1! x_2! \cdots x_k!} u_1^{x_1} \cdots u_k^{x_k}.$$

よって、負の多項分布の場合と同様、Pooling incomplete samples が「多項分布」の部分で行われていれば Total = Between + Within が成り立つ。

文献 [1] Asano, C. (1965). On estimating multinomial probabilities by pooling incomplete samples, *Annals of the Institute of Statistical Mathematics* **17**, 1-13. [2] 布能 英一郎 (2015). 負の多項分布における Kullback 情報量の直和分解—— Pooling incomplete samples の場合を含めた考察—— 京都大学数理解析研究所講究録 No 1954, 90-103. [3] Johnson, N.L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions.* Wiley. [4] Kullback, S. (1968). *Information Theory and Statistics.* Revised versuin Dover.

判別関数の誤分類確率と判別係数の 95%信頼区間

成蹊大学 経済学部　新村秀一

判別分析には，次の 4 つの問題がある．

1)　誤分類数最小化（Minimum Number of Misclassifications, **MNM**）基準による最適線形判別関数（**改定 IP-OLDF**）以外の判別関数は，判別超平面上のケースを正しく判別できないので，正しい誤分類数が求まらない．

2)　ハードマージン最大化 SVM（**H-SVM**）と**改定 IP-OLDF** 以外の判別関数は，線形分離可能なデータ（**MNM=0**）を認識できないことが多い．そして公表している実証研究で Fisher の線形判別関数（Fisher の LDF）は，線形分離可能な 18 個の分析結果で，誤分類確率の範囲が[0.17,0.23]であることが分かった．すなわち，過去の医学診断等で例え誤分類確率が 23%であっても，実際は線形分離可能な可能性があり過去の重要な判別分析を用いた研究を見直す必要がある．

3)　変数値が一定値をとる場合，Fisher の LDF と 2 次判別関数（QDF）は分散共分散行列が計算できないので分析からこれらの変数を省く統計ソフトが多い．一般化逆行列を用いればこれに対応できるが，一方の群に属する変数値が一定値をとる場合に QDF と正則化判別分析で深刻な問題を起こすことが分かった．

4)　判別分析を重回帰分析と同じ推測統計学的手法と間違って考えている人が多い．しかし，重回帰分析のように判別係数や誤分類確率の 95%信頼区間がなく，重回帰分析のような洗練されたモデル選択法がない．

本報告では、「小標本のための k-重交差検証法」を用いて、誤分類確率と判別係数の 95%信頼区間を検討した。検討した手法は、改定 IP-OLDF を Fisher の LDF，ロジスティック回帰，H-SVM（MNM=0 のデータのみ），ソフトマージン最大化 SVM（S-SVM），改定 LP-OLDF，改定 IPLP-OLDF の 7 個の線形判別関数を用いて，学習標本と検証標本の平均誤分類確率と判別係数の 95%信頼区間を検討した．検証に用いたデータの一例として、一番結果のはっきりする 18 種類の統計入門の 10 卓 100 問の合格水準を 10%，50%，90%の 3 水準による合否判定で行ったが、その一例を紹介する．

表 3 は、信頼区間の変わりに学習標本と検証標本の誤分類確率の範囲を示す。

Table 3. The ranges of error rates.

	T1, T2, T3, T4		T1, T2, T4		T2, T3, T4		T2, T4	
	MIN	MAX	MIN	MAX	MIN	MAX	MIN	MAX
RIP[*2]	0	0	0	0	0	0	0	0
	0	4.84	0	3.23	0	4.84	0	4.84
SVM1	0	0	0	2.42	0	0.81	0	3.23
	0	4.84	0.81	3.23	0	4.84	0	4.84
LP	0	0	0	0	0	0	0	0
	0	4.84	0	3.23	0	4.84	0	3.23
LDF	4.03	16.13	4.03	16.94	4.03	16.13	4.03	17.74
	6.40	12.82	8.01	12.82	8.00	13.60	8.00	13.60

Upper: training sample, Lower: validation sample.

*2: RIP, HSVM, SVM4, IPLP, and Logistic are the same.

表 5 は、判別係数の 95%信頼区間である。

Table 5. The 95% C.I. of six LDFs.

RIP	T1	T2	T3	T4	c
97.5	0.384	1.413	1.396	0.438	-6.253

Median	0.149	0.654	0.405	0.314	-13.142
2.5	-0.108	0.242	-0.018	0.171	-18.378
HSVM	T1	T2	T3	T4	c
97.5	0.483	1.170	0.761	0.655	-6.581
Median	0.256	0.623	0.443	0.369	-13.994
2.5	-0.031	0.218	-0.025	0.193	-22.351
SVM4	T1	T2	T3	T4	c
97.5	0.483	1.170	0.761	0.655	-6.582
Median	0.256	0.623	0.443	0.369	-13.989
2.5	-0.028	0.219	-0.025	0.193	-22.351
SVM1	T1	T2	T3	T4	c
97.5	0.483	1.170	0.761	0.655	-6.582
Median	0.256	0.623	0.443	0.370	-13.994
2.5	-0.031	0.218	-0.025	0.193	-22.351
IPLP	T1	T2	T3	T4	c
97.5	0.579	1.366	1.462	0.655	-6.029
Median	0.171	0.651	0.418	0.341	-13.252
2.5	-0.123	0.058	-0.048	0.179	-22.689
LP	T1	T2	T3	T4	c
97.5	0.412	1.677	1.428	0.451	-6.047
Median	0.149	0.705	0.405	0.330	-13.060
2.5	-0.158	0.177	-0.036	0.172	-18.773

共同プロジェクトの提案

判別分析にかかわる研究を見直すために，学会あるいは医学部と共同研究を立ち上げることを提案したい．筆者は，判別分析にかかわる計算を引き受ける，あるいはそれに関する技術を提供することで協力したいと考えている．このようなテーマを持っておられる関係者を知っておられる場合は，ぜひ紹介してください．

参考文献

12） J. Sall（新村秀一訳）: SAS による回帰分析の実践，朝倉書店，東京,1986.

13）J.P. Sall, L. Creighton, & A. Lehman（新村秀一監修）: JMP を用いた統計およびデータ分析入門（第 3 版）. SAS Institute Japan ㈱, 2004.

14）新村秀一，高森寛，実践数理計画法．朝倉書店,1992.

15）L. Schrage: Optimizer Modeling with LINGO. LINDO Systems Inc, 2003.

21）新村秀一: 数理計画法を用いた最適線形判別関数．計算機統計学．**11**/2,89-101,1997.

24）新村秀一: JMP による統計学とっておき勉強法，講談社，2004.

27）新村秀一: Excel と LINGO で学ぶ数理計画法，丸善，2007.

30）新村秀一: 最適線形判別関数．日科技連出版社，2010.

31）新村秀一: 数理計画法による問題解決法．日科技連出版社，2011.

33）新村秀一: SAS/JMP との歩み，SAS Technical Report，**13**/16，2012.

34）S. Shinmura: End of Discriminant Functions based on Variance Covariance Matrices. ICORER, 5-14, 2014a.

35）S. Shinmura, “Improvement of CPU time of Linear Discriminant Functions based on MNM criterion by IP,” Statistics, Optimization and Information Computing, vol. 2, June 2014, pp 114-129.

36）S. Shinmura(2014). Comparison of Linear Discriminant Functions by K-fold Cross Validation. Data Analytic 2014, 1-6, 2014.

Statistical Inference for Ergodic Point Processes and Applications to Limit Order Book*

Simon Clinet

PhD student under the supervision of Pr. Nakahiro Yoshida

Graduate School of Mathematical Sciences, University of Tokyo

CREST, Japan Science and Technology Agency

Toyama Symposium

October 2, 2015

Most financial transactions take place nowadays in electronic markets. Participating to continuous-time double auctions, agents can freely send buying or selling orders at different prices that are automatically matched according precise rules. As this matching process is rather complex and the orders sent by market participants are asynchronous, they are centralized in an Electronic Limit Order Book (also denoted LOB), waiting to be executed according to their price and time priority. A LOB is thus a multidimensional queuing system, each dimension representing a price level, and each queue containing the waiting orders that have not been executed yet, sorted by their arrival time. Agents can then interact with this dynamical system via three elementary mechanisms. They may submit a buying (resp. selling) limit order that will increase the size of one queue on the bid (resp. ask) side of the LOB. They also may send a buying (resp. selling) market order that will immediately consume the corresponding liquidity at the best available price. Finally they can submit cancellations orders to remove one of their latent limit order in the LOB.

Since a LOB is mechanically driven by the orders that are submitted through time, many authors choose to see a LOB through the stochastic structure of the interarrival times between two events. In other words, a Limit Order Book is often described as a high-dimensional point process, whose components are integer measures counting the waiting times between two orders of the same type, and at the same price level. In a parametric context, estimating the parameter θ^* based on the observations is thus a very crucial issue that may take place in two distinct asymptotics. As, at least for liquid stocks, a tremendously large number of events happen during short periods of times, the heavy traffic limit seems to be a good way to construct consistent estimators. In [5], a sequence of multivariate point processes is thus assumed to be observable on a finite time window. Under suitable assumptions on the sequence of stochastic intensities itself, it is shown that even in this non-ergodic context it is possible to conduct the quasi likelihood analysis procedure (QLA for short).

On the other hand, in this work, we are interested in the long run characteristics of the LOB seen as a point process. as the time parameter T tends to infinity, assuming that the LOB satisfies suitable ergodicity assumptions, we aim at taking advantage of this regularity in order to derive the asymptotic properties of the QMLE and the QBE. This problematic is of course not new since the consistency and the asymptotic normality

*This work was in part supported by CREST Japan Science and Technology Agency; Japan Society for the Promotion of Science Grants-in-Aid for Scientific Research No. 24340015 (Scientific Research), No. 26540011 (Challenging Exploratory Research); NS Solutions Corporation; and by a Cooperative Research Program of the Institute of Statistical Mathematics.

of the maximum likelihood estimator for ergodic stationary point processes was shown a few decades ago in [4] and [6]. Furthermore, maximum likelihood estimations have also been empirically conducted for the above-mentioned models, but the fact that the point process is ergodic is sometimes unclear. In this article we thus give general ergodicity and regularity assumptions for point processes under which all the results from the general QLA can be derived. In particular, we do not necessarily require the stationarity of the point process.

More precisely, for a given a multivariate point process $N_t = (N_t^\alpha)_{\alpha \in \mathbf{I}}$, $\mathbf{I} = \{1, ..d\}$, and some finite dimensional compact space $\Theta \subset \mathbb{R}^n$, we consider the family of increasing processes $\Lambda_t(\theta) = \int_0^t \lambda(s, \theta) ds$, and we assume that there exists $\theta^* \in \Theta$ such that $\lambda(t, \theta^*)$ is the $\mathcal{F}_t$-intensity of N_t. Under the key assumption,

[$A3^{'}$] **Ergodicity.** There exists a mapping $\pi : C_\uparrow(\mathbb{R}^4, \mathbb{R}) \times \Theta \to \mathbb{R}$ and there exists $0 < \gamma < \frac{1}{2}$ such that for any $(\psi, \theta) \in C_\uparrow(\mathbb{R}^4, \mathbb{R}) \times \Theta$ the following convergence holds :

$$\sup_{\theta \in \Theta} T^\gamma \left\| \frac{1}{T} \int_0^T \psi(\lambda(s, \theta^*), \lambda(s, \theta), \partial_\theta \lambda(s, \theta), \partial_\theta^2 \lambda(s, \theta)) ds - \pi(\psi, \theta) \right\|_p \to 0 \tag{1}$$

for a precise class of functions $C_\uparrow(\mathbb{R}^4, \mathbb{R})$,

we show that under [$A3^{'}$] and some other regularity assumptions, we have the following result :

Theorem 0.1. *Under [$A1^{'}$]-[$A4^{'}$], the following result holds:*
If $\hat{\theta}_T$ is the QMLE and $\tilde{\theta}_T$ the QBE, there exists Γ such that we have :

$$\mathbb{E}\left[f(\sqrt{T}(\hat{\theta}_T - \theta^*)) \right] \to \mathbb{E}[f(\Gamma^{-\frac{1}{2}} \mathcal{N})]$$

$$\mathbb{E}\left[f(\sqrt{T}(\tilde{\theta}_T - \theta^*)) \right] \to \mathbb{E}[f(\Gamma^{-\frac{1}{2}} \mathcal{N})]$$

for any continuous f with polynomial growth.

We then check that some classical models from the literarure satisfy this ergodicity condition. More precisely, we show that V-geometric ergodic LOB [1, 3], Cox point processes depending on an ergodic Markov process [7], and finally exponential Hawkes-driven LOB [2], are all in the class of models for which the QLA applies.

References

[1] Abergel, F., Jedidi, A.: A mathematical approach to order book modeling. International Journal of Theoretical and Applied Finance **16**(5) (2013)

[2] Abergel, F., Jedidi, A.: Long time behaviour of a hawkes process-based limit order book. Preprint (2015)

[3] Huang, W., Lehalle, C., Rosenbaum, M.: Simulating and analyzing order book data: The queue-reactive model. SIAM Journal on Financial Mathematics **4**(1) (2014)

[4] Ogata, Y.: The asymptotic behaviour of maximum likelihood estimators for stationary point processes. Annals of the Institute of Statistical Mathematics **30**(A), 243–261 (1978)

[5] Ogihara, T., Yoshida, N.: Quasi likelihood analysis for point process regression models. preprint (2015)

[6] Puri, M.L., Tuan, P.D.: Maximum likelihood estimation for stationary point process. Proceedings of the National Academy of Sciences **83**, 541–545 (1986)

[7] Rosenbaum, M., Delattre, S., Robert, C.Y.: Estimating the efficient price from the order flow : a brownian cox process approach. arXiv:1301.3114 (1971)

The locally parametric model: a new class of models in high frequency data

Yoann Potiron

November 2, 2015

Abstract

This paper proposes mixed parametric and nonparametric statistical techniques for the analysis of high frequency data. It gives a general model, which can be discrete or continuous in time depending on the point-of-view. This model can be seen as a parametric model which allows its multidimensional parameter to follow a local martingale. As such, we call it the *locally parametric model* (LPM). The quantity of interest is defined as the *uniformly weighted value over time* (UWV) of the (discrete or continuous) parameter process. We provide estimators of UWV and conditions under which we can show the consistency and the corresponding central limit theorem. Those estimators are based on estimators of the parametric model when parameters are fixed. Since the estimator is obtained by chopping the data into small blocks, estimating the parameter on each block pretending it is constant locally and taking a weighted mean of the estimates on each block, we call it the *locally parametric quasi-estimator* (LPQE). We show that under conditions, some discrete standard time series models of the literature (for instance ARMA or GARCH models with MLE estimator) as well as continuous semiparametric models (for example a semimartingale asset price model with IID noise component in the observations) of the high-frequency financial econometrics literature belongs to the LPM class of models. This paper thus builds a bridge between various perspectives, parametric, semiparametric and nonparametric as well as discrete and continuous in time models. In addition, statistics to test whether the parameter's constancy hypothesis is true are provided. We also discuss model selection and provide statistics to test for nested models: as an example, this allows us to test if there is noise in observations. Based respectively on the estimate of UWV, we give a new input to use in the prediction model. Finally, an empirical study on S & P 500 daily returns, using ARMA and models is carried out. It shows that the parameters are not constant over time for both models and that we obtain better statistical inference using the new prediction's input of the model.

Modeling dynamics is very important in various fields, such as finance, economics, physics, environmental engineering, geology or even sociology. Parametric time dependent models are tools meant to deal with one type of dynamics, the temporal evolution of systems. There has been an explosion of research in the area in the last decades. We can identify two main reasons why parametric models are very attractive and popular, both for researchers and practitioners. First, by estimating an underlying (possibly multidimensional) parameter, they provide crucial information on the mechanisms of the system of interest. As an example, the fitted parameter of *autoregressive moving average* (ARMA) models (Whittle (1951)) give us insight on the correlation structure of the observations. Also, parametric models usually allow for inference such as prediction of future observations together with confidence intervals, as a function of the data. In particular, if we choose an adequate model, we can predict tomorrow's temperature.

By definition, parametric approaches come with the strong assumption that there exists an underlying parameter, who drives the structure of the observations, and which is fixed over time. In practice, the parametric model user usually tries different types of models, or has a specific class of models in mind, and she fits the models to the data. It means that she estimates the parameter of the model with the observations. Nonetheless, as time goes by, the structure driving the observations is most likely evolving as well. Thus, questions about the constancy of the parameter, that would stay the same through thick and thin, are to be raised. To corroborate this natural skepticism, it can even be the case that empirical work strongly suggests that the assumption of constancy is too restrictive. To acknowledge the issue, one has to build an extended model, that can be either parametric but typically with some more parameters, semiparametric or nonparametric.

The extremogram and the cross-extremogram for a bivariate GARCH(1,1) process

Muneya Matsui 松井 宗也

Firstly we define a measure for exteremal serial dependence in a bivariate series. Our main tool is the extremogram and cross-extremogram of a bivariate sequence $(\mathbf{X}_t) = (X_{1t}, X_{2t})'$ in standardized form such that they assume values in $[0,1]$:

$$\begin{pmatrix} \rho_{11}(h) & \rho_{12}(h) \\ \rho_{21}(h) & \rho_{22}(h) \end{pmatrix}, \qquad h = 0, 1, 2, \ldots,$$

where

$$\begin{pmatrix} \rho_{11}(h) \\ \rho_{22}(h) \\ \rho_{12}(h) \\ \rho_{21}(h) \end{pmatrix} = \lim_{x\to\infty} \begin{pmatrix} \mathbb{P}(X_{1,h} \in x\,A \mid X_{1,0} \in x\,A) \\ \mathbb{P}(X_{2,h} \in x\,B \mid X_{2,0} \in x\,B) \\ \mathbb{P}(X_{2,h} \in x\,B \mid X_{1,0} \in x\,A) \\ \mathbb{P}(X_{1,h} \in x\,A \mid X_{2,0} \in x\,B) \end{pmatrix}. \tag{0.1}$$

Here A, B are sets bounded away from zero and we assume that these limits exist. Typically, we choose intervals $(1,\infty)$, $(-\infty,-1)$ for A, B and we also suppress the dependence on A, B in the ρ_{ij}-notation. The limits $\rho_{ij}(h)$ in (0.1) do not automatically exist. A convenient theoretical assumption for their existence is the condition of *regular variation of the time series* $(\mathbf{X}_t)$ (intuitively, power law behavior).

Next, we focus on a bivariate GARCH$(1,1)$ model which is common proxy to analyze financial data, and clarify the component-wise tail behavior of the process, i.e. we show that the component of a bivariate GARCH$(1,1)$ process may exhibit regular variation (power law behavior). With the result we will analyze the extremogram and the cross-extremogram for the bivariate GARCH$(1,1)$ process. Recall that $(\mathbf{X}_t)$ has the following structure:

$$\mathbf{X}_t = \Sigma_t\,\mathbf{Z}_t\,, \qquad t \in \mathbb{Z}\,, \tag{0.2}$$

where $(\mathbf{Z}_t) = (Z_{1,t}, Z_{2,t})'$ constitutes an iid bivariate noise sequence and

$$\Sigma_t = \mathrm{diag}(\sigma_{1,t}, \sigma_{2,t})\,, \quad t \in \mathbb{Z}\,,$$

where $\sigma_{i,t}$ is the (non-negative) volatility of $X_{i,t}$. The model has the following specification by the stochastic recurrence equating (SRE) by $(\sigma^2_{1,t}, \sigma^2_{2,t})'$:

$$\begin{aligned} \begin{pmatrix} \sigma^2_{1,t} \\ \sigma^2_{2,t} \end{pmatrix} &= \begin{pmatrix} \alpha_{01} \\ \alpha_{02} \end{pmatrix} + \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} \begin{pmatrix} X^2_{1,t-1} \\ X^2_{2,t-1} \end{pmatrix} + \begin{pmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{pmatrix} \begin{pmatrix} \sigma^2_{1,t-1} \\ \sigma^2_{2,t-1} \end{pmatrix} \\ &= \begin{pmatrix} \alpha_{01} \\ \alpha_{02} \end{pmatrix} + \begin{pmatrix} \alpha_{11}Z^2_{1,t-1} + \beta_{11} & \alpha_{12}Z^2_{2,t-1} + \beta_{12} \\ \alpha_{21}Z^2_{1,t-1} + \beta_{21} & \alpha_{22}Z^2_{2,t-1} + \beta_{22} \end{pmatrix} \begin{pmatrix} \sigma^2_{1,t-1} \\ \sigma^2_{2,t-1} \end{pmatrix}. \end{aligned}$$

Writing $\mathbf{W}_t = (\sigma^2_{1,t}\,, \sigma^2_{2,t})'$, we obtain the stochastic recurrence equation

$$\mathbf{W}_t = \mathbf{A}_t\,\mathbf{W}_{t-1} + \mathbf{B}_t\,, \qquad t \in \mathbb{Z}\,. \tag{0.3}$$

Therefore, we can invoke theoretical results of SRE, which yields the following proposition.

Proposition 0.1. *Consider the bivariate* GARCH$(1,1)$ *model and assume the following conditions:*
(1) $\gamma = \lim_{n\to\infty} n^{-1} \log \|\mathbf{A}_1 \cdots \mathbf{A}_n\| < 0$.
(2) $\mathbf{Z}_0$ *has Lebesgue density in* $\mathbb{R}^2$.
(3) *There exists* $p > 0$ *such that*

$$\mathbb{E}[|\mathbf{Z}_0|^{2p} \log^+ |\mathbf{Z}_0|] < \infty \quad \textit{and} \quad \mathbb{E}\big[\min_{i=1,2} \big(\sum_{j=1}^{2} (\alpha_{ij} Z^2_{j,0} + \beta_{ij})\big)^p\big] \geq 2^{p/2}\,. \tag{0.4}$$

(4) *All entries of* $\mathbf{A}_0$ *are positive a.s.,* $\alpha_{0i} > 0$, $i = 1, 2$, *and not all values* α_{ij}, $1 \le i, j \le 2$, *vanish. Then there exists a unique* $\alpha \in (0, 2\,p]$ *such that*

$$0 = \lim_{n\to\infty} n^{-1} \log \mathbb{E}\big[\|\mathbf{A}_1 \cdots \mathbf{A}_n\|^{\alpha/2}\big]\,, \tag{0.5}$$

there exists a strictly stationary causal non-zero solution $(\mathbf{X}_t)$ *to* (0.2) *with specification* (0.3) *and* $(\mathbf{X}_t)$ *is regularly varying with index* α. *In particular, for every* $n \ge 1$, *there exists a non-null Radon measure* μ_n *on* $\overline{\mathbb{R}}^{2n}\backslash\{\mathbf{0}\}$, $\overline{R} = \{-\infty, \infty\} \cup \mathbb{R}$, *such that*

$$x^\alpha\, \mathbb{P}\big(x^{-1}(\mathbf{X}_1, \ldots, \mathbf{X}_n) \in \cdot\big) \xrightarrow{v} \mu_n(\cdot)\,, \qquad x \to \infty\,.$$

Here $\xrightarrow{v}$ *denotes vague convergence and the limit measures have the property* $\mu_n(t\cdot) = t^{-\alpha}\mu_n(\cdot)$, $t > 0$.

Now, with this proposition we analyze the asymptotics of extremogram and the cross-extremogram for the GARCH(1, 1) process. Since the tail behavior of $(\mathbf{X}_t)$ and that of $\mathbf{W}_t = (\sigma_{1,t}^2\,, \sigma_{2,t}^2)'$ are proportional, we restrict ourselves to the σ-sequences. By induction $\mathbf{W}_t = \mathbf{\Pi}_t\,\mathbf{W}_0 + \mathbf{R}_t$, where $\mathbf{\Pi}_t = \mathbf{A}_t \cdots \mathbf{A}_1$, $\mathbf{R}_t = \sum_{i=1}^{t-1} \mathbf{A}_t \cdots \mathbf{A}_{t-i+1}\mathbf{B}_{t-i} + \mathbf{B}_t$ for $t \ge 1$, and all vectors are interpreted as column vectors. With this interpretation we write

$$(\mathbf{W}_1, \ldots, \mathbf{W}_t) = (\mathbf{\Pi}_1, \ldots, \mathbf{\Pi}_t)\,\mathbf{W}_0 + (\mathbf{R}_1, \ldots, \mathbf{R}_t)\,, \qquad t \ge 1\,, \tag{0.6}$$

where $(\mathbf{\Pi}_1, \ldots, \mathbf{\Pi}_t)$, $(\mathbf{R}_1, \ldots, \mathbf{R}_t)$ have moment of order $\alpha/2$ with respect to the corresponding matrix norms and are independent of $\mathbf{W}_0$. We assume the conditions of Proposition 0.1; in this case both components $\sigma_{i,t}^2$, $i = 1, 2$, of the vector $\mathbf{W}_t$ in (0.3) have the same tail index. Using relation (0.6), we see that

$$\begin{aligned}
\rho_{ij}(h) &= \lim_{x\to\infty} \mathbb{P}(\sigma_{j,h}^2 > x \mid \sigma_{i,0}^2 > x) = \lim_{x\to\infty} \frac{\mathbb{P}(\sigma_{j,h}^2 > x, \sigma_{i,0}^2 > x)}{\mathbb{P}(\sigma_{i,0}^2 > x)} \\
&\le \lim_{x\to\infty} \frac{\mathbb{P}(|\mathbf{W}_h| > x, |\mathbf{W}_0| > x)}{\mathbb{P}(|\mathbf{W}_0| > x)} \times \frac{\mathbb{P}(|\mathbf{W}_0| > x)}{\mathbb{P}(\sigma_{i,0}^2 > x)}\,.
\end{aligned}$$

The limit of the latter ratio converges to a constant by virtue of regular variation. Thus the extremograms ρ_{ij} are bounded by the extremogram $\rho_{|\mathbf{W}|}$ of $(|\mathbf{W}_t|)$ times this constant. However, (0.6) and the independence of $\mathbf{W}_0$ and $\mathbf{R}_h$ imply that for $p < \alpha/2$ and $h \ge 1$

$$\begin{aligned}
\rho_{|\mathbf{W}|}(h) &= \lim_{x\to\infty} \frac{\mathbb{P}(|\mathbf{W}_h| > x, |\mathbf{W}_0| > x)}{\mathbb{P}(|\mathbf{W}_0| > x)} \\
&\le \limsup_{x\to\infty} \frac{\mathbb{P}(\|\mathbf{\Pi}_h\|\,|\mathbf{W}_0| > x/2\,, |\mathbf{W}_0| > x)}{\mathbb{P}(|\mathbf{W}_0| > x)} + \lim_{x\to\infty} \mathbb{P}(|\mathbf{R}_h| > x/2) \\
&= \text{const}\,\mathbb{E}[\min(1, \|\mathbf{\Pi}_{\rm h}\|^{\alpha/2})] \le \text{const}\,\mathbb{E}[\min(1, \|\mathbf{\Pi}_{\rm h}\|^{\rm p})] \le \text{const}\,\mathbb{E}[\|\mathbf{\Pi}_{\rm h}\|^{\rm p}].
\end{aligned}$$

The right-hand side converges to zero at an exponential rate in view of $\mathbb{E}[\|\mathbf{\Pi}_{h_0}\|^p] < 1$ for a sufficiently large h_0. As a result, $\rho_{|\mathbf{W}|}(h)$ is shown to converge to zero exponentially fast.

We apply the theory to 5-minute return data of stock prices and foreign exchange rates, i.e. we judge the fit of a bivariate GARCH(1, 1) model by considering the sample extremogram and cross-extremogram of the residuals. The results are in agreement with the iid hypothesis of the two-dimensional innovations sequence. The cross-extremograms at lag zero have a value significantly distinct from zero. This fact points at some strong extremal dependence of the components of the innovations.

References

[1] Davis, R.A. and Mikosch, T. (2009) The extremogram: a correlogram for extreme events. *Bernoulli* **15**, 977-1009.

[2] Matsui, M. and Mikosch, T. (2015) The extremogram and the cross-extremogram for a bivariate GARCH(1,1) process. J. Appl. Prob. (to appear), see also ArXiv:1505.05385.

日経225平均株価指数の日次収益率分析におけるジャンプ拡散過程モデルの同定とその低頻度で振幅の大きなジャンプ時点推定への応用について

石田真之（東京都市大学大学院工学研究科）、金川秀也（東京都市大学 共通教育部）、

日経225平均株価指数の日次収益率に対してブラック・ショールズモデルと複合ポアソン過程によって構成されたジャンプ拡散過程によってモデリングを行った場合に、複合ポアソン過程から発生した日次収益率のジャンプ時点を推定する. 特に一日の終値による日次収益率は完全な離散データであるため、これらの時系列データから連続部分とジャンプ部分を分離することは容易ではない. 本報告における複合ポアソン過程を同定する手法は、ヒストリカル・ボラティリティを基準として振幅の大きなジャンプを判別するため、株価の連続部分のモデル形（ブラック・ショールズモデル）をフルに使用しておらず、株価データから得られるトレンドとボラティリティの推定値しか必要としない. ヒストリカル・ボラティリティの観測期間の選び方で大きく分析結果が異なることから、株価を観測する単位期間の長さとヒストリカル・ボラティリティの観測期間の長さの間に適切な比率が存在することを実証する. またジャンプ時点を抽出する問題は株価分析の応用上極めて重要であり、提唱された手法は株価モデルへの頑健性が高いことから、応用面から有効であると期待される.　株価モデルとしてボラティリティ変動モデルである、次に定義されるジャンプ拡散過程を用いる.

株価のトレンドを μ_t、ボラティリティを σ_t とする.

$$S(t) = S(t, \mu_t, \sigma_t) = \tilde{S}(t, \tilde{\mu}_t, \tilde{\sigma}_t) + \hat{S}(t, \hat{\mu}_t, \hat{\sigma}_t), \quad 0 \leq t \leq T. \tag{1}$$

ただし、$\tilde{S}(t, \tilde{\mu}_t, \tilde{\sigma}_t)$ はトレンド $\tilde{\mu}_t$, ボラティリティー $\tilde{\sigma}_t$ である道が連続な確率過程、$\hat{S}(t, \hat{\mu}_t, \hat{\sigma}_t)$ はジャンプ過程とする. $S(t)$ は確率微分方程式

$$dS(t) = dS(\mu_t, \sigma_t, t) = S(t)\, dX(t) + S(t)\, dZ(t), \quad 0 \leq t \leq T \tag{2}$$

を満たすと仮定する. ただし, $X(t)$ は連続確率過程、$Z(t)$ は複合ポアソン過程であり、互いに独立とする. また $B(t)$ を標準ブラウン運動とする.　$X(t)$ に対して確率微分方程式

$$dX(t) = \tilde{\mu}_t dt + \tilde{\sigma}_t dB(t), \quad 0 \leq t \leq T \tag{3}$$

が成り立つとき、$S(t)\, dX(t)$ は連続なブラック・ショールズモデルを表し、また $S(t)$ は複合ポアソン過程をジャンプ部分とするジャンプ拡散過程である.

日経225平均株価指数に対してジャンプ拡散過程によるモデリングを行う場合に、株価データから不連続な複合ポアソン過程の部分を抽出する手法について考察する。一日の終値や

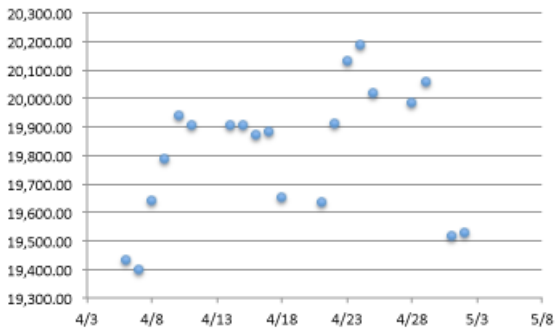


図 1: 日経 225 平均株価指数, 2015.4.7∼5.2

その日次収益率は離散データであるため、これらの時系列データから連続部分とジャンプ部分を分離することは容易ではない. 図 1 は日経 225 平均株価の 2015 年 4 月 7 日から 2015 年 5 月 2 日の一日の終値を表す実データである.

ジャンプ拡散過程から通常の変動による連続過程と不連続過程を分離することは、株価分析上で非常に有効である. 特に振幅の大きな不連続部分では、株価の急激で大きな変動を助長する情報が市場に流れた場合や、株価の上昇や下降が長期間続き限界に達して急激な売買が行われた場合が想定され、応用上特に注目される. 本報告ではこの点に注目して、数理統計学の見地から低頻度で振幅の大きなジャンプが生じた時点を推定する手法を提案する.

株価の実証分析において Bell and Torous (1983, 1985) は Merton モデルの推定を行い、株価の変動におけるジャンプの存在を示した. それ以後、株価データをジャンプ拡散過程によってモデル化してパラメータの推定を行う手法について多くの研究がなされてきた. 本報告と同様に振幅の大きなジャンプを検出することを主目的とする研究として、Kwakernaak (1980) や飯野・尾崎 (1999) がある.

参考文献

[1] Ball, C. A. and Torous, W. N.:“A Simplified Jump Process for Common Stock Returns,” *Journal of Financial and Quantitative Analysis*, Vol.18, pp.53–65 (1983)

[2] Kwakernaak, H.:“Estimation of Pulse Heights and Arrival Times,” *Automatica*, Vol.16, pp.367–377 (1980)

[3] 飯野三徳、尾崎統: “ジャンプ拡散過程モデルによる非ガウス時系列のフィルタリングと予測”, 統計数理, Vol.47, No.2, pp.327–342 (1999)

3次元分割表における
1要因対2要因の独立性検定による改良変換統計量

小部 敬純 *
種市 信裕 †
関谷 祐里 ‡

1 はじめに

本講演では，多項分布モデルの 3 次元分割表における row カテゴリと column, layer カテゴリとが独立であるという帰無仮説のもとでの ϕ-ダイバージェンス統計量 C_ϕ の分布を考察する．一般に，(3 次元) 分割表のもとでの C_ϕ は標本が十分大きいとき χ^2 分布を極限分布を持つことはすでに知られており，C_ϕ の特別な場合として，パワーダイバージェンス統計量 R^a も含まれる．これらの統計量についての関心事の一つに漸近展開による近似があげられる．

多項分布の適合度検定において，Yarnold (1972) はピアソンの χ^2 統計量における漸近展開を導き，数値実験から漸近展開に基づく近似の正確さを示した．その後， Siotani and Fujikoshi (1984) は対数尤度比統計量と Freeman-Tukey 統計量を考察し，パワーダイバージェンス統計量は Read (1984) によって示されている．また，ϕ-ダイバージェンス統計量についての漸近近似は Menéndez *et al.* (1997) が導いている．そして，Taneichi and Sekiya (2007) は 2 次元分割表での独立性検定における改良変換統計量について ϕ-ダイバージェンス統計量を用いて論じた．

本講演においては，C_ϕ の分布を連続項と仮定し，帰無仮説のもとでの C_ϕ の分布における多変量 Edgeworth 展開に基づく近似式の導出を考察する．まず，row カテゴリと column, layer カテゴリとが独立であるモデルの C_ϕ とその特別な場合である R^a を紹介し，帰無仮説のもとでの局所 Edgeworth 近似について説明する．そして，C_ϕ の分布の多変量 Edgeworth 展開についての考察をしている．さらに，バートレット修正と，改良変換統計量について概説し，多変量 Edgeworth 展開に基づいた新しい変換統計量を提案する．

キーワード
カイ二乗極限分布 ; 分割表 ; 改良変換 ; 1 要因対 2 要因の独立性検定 ; ϕ-ダイバージェンス ; パワーダイバージェンス

2 独立性のモデル

$r \times s \times t$ 分割表において，セル確率を p_{ijk} $(i = 1, \ldots, r;\ j = 1, \ldots, s;\ k = 1, \ldots, t)$ とする．また，多項分布モデルについて考えるので，

$$\boldsymbol{X} \sim \mathrm{Mult}_{rst}(n; \boldsymbol{p})$$

とする．また帰無仮説を

$$H_0 : p_{ijk} = p_{i\cdot\cdot} p_{\cdot jk} \tag{2.1}$$

*鹿児島大学 大学院 理工学研究科 システム情報科学専攻 D2
†鹿児島大学 大学院 理工学研究科 数理情報科学専攻
‡北海道教育大学 釧路校

とする．H_0 のもとでの，p_{ijk} の最尤推定量は，$\hat{p}_{ijk} = \hat{p}_{i\cdot\cdot}\hat{p}_{\cdot jk}$ と与えらる．

ϕ-ダイバージェンス統計量 C_ϕ は，

$$C_\phi = 2n \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t} \hat{p}_{i\cdot\cdot}\hat{p}_{\cdot jk}\phi\left(\frac{\hat{p}_{ijk}}{\hat{p}_{i\cdot\cdot}\hat{p}_{\cdot jk}}\right)$$

と与えられている．ただし，$\phi(t)$ は，$t > 0$ において，実凸関数であり，$\phi(1) = \phi'(1) = 0, \phi''(1) = 1$ を満たす．ここで，凸関数を

$$\phi_a(t) = \begin{cases} \{a(a+1)\}^{-1}\{t^{a+1} - t + a(1-t)\} & (a \neq 0, -1), \\ t\log t + 1 - t & (a = 0), \\ -\log t - 1 + t & (a = -1), \end{cases}$$

とすることで，C_ϕ は

$$C_{\phi_a} \equiv R^a = 2n \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t} I^a(\hat{p}_{ijk},\ \hat{p}_{i\cdot\cdot}\hat{p}_{\cdot jk}), \tag{2.2}$$

となる．ただし，R^a はパワーダイバージェンス統計量である．このとき，

$$I^a(e,f) = \begin{cases} \dfrac{1}{a(a+1)} e\left\{\left(\dfrac{e}{f}\right)^a - 1\right\} & (a \neq 0, -1), \\ e\log\dfrac{e}{f} & (a = 0), \\ f\log\dfrac{f}{e} & (a = -1). \end{cases}$$

パワーダイバージェンス統計量を用いることで，対数尤度比統計量を R^0，ピアソンの χ^2 統計量を R^1 と表現することができる．そして， $R^{2/3}$ は Cressie and Read (1984) によって推奨された統計量である．また， H_0 のもとで， C_ϕ は自由度 $(r-1)(st-1)$ の χ^2 極限分布を持つ．

参考文献

[1] N. Cressie, T. R. C. Read (1984). Multinomial goodness-of-fit tests, *J. Roy. Statist. Soc. B* **46**, 440-464.

[2] M. L. Menéndez, J. A. Pardo, L. Pardo, M. C. Pardo (1997). Asymptotic approximations for the distributions of the (h,ϕ)-divergence goodness-of-fit statistics: application to Renyi's statistics, *Kybernetes* **26 (4)**, 442-452.

[3] T. R. C. Read (1984). Closer asymptotic approximations for the distributions of the power divergence goodness-of fit statistics, *Ann. Inst. Statist. Math.* **36**, 59-69.

[4] M. Siotani,Y. Fujikoshi (1984). Asymptotic approximations for the distributions of multinomial goodness-of-fit statistics, *Hiroshima Math. J.* **14**, 115-124.

[5] N. Taneichi, Y. Sekiya (2007). Improved transformed statistics for the test of independence in $r \times s$ contingency tables, *J. Multivariate Anal.* **98**, 1630-1657.

[6] J. K. Yarnold (1972). Asymptotic approximations for the probability that a sum of lattice random vectors lies in a convex set, *Ann. Math. Statist.* **43**, 1566-1580.

高次元枠組みにおける共分散構造に関する検定について

専修大・経営　　西山　貴弘
東京理科大・理・院　山田　雄紀
大阪府立大・工　　兵頭　昌

本報告では共分散構造に関する検定問題について議論を行った. 一般に, $p < N$ の場合, 共分散構造の検定問題に対して尤度比検定が用いられるが, $p \geq N$ の場合は標本共分散行列や標本相関行列が退化してしまうため尤度比検定統計量を構成することが出来ないという問題が生じる. そこで近年, 高次元データに対して球面構造や対角構造を持つかどうかの検定や, 共分散行列の同等性検定問題などについて多くの研究がされている（Schott (2005), Chen, Zhang and Zhong (2010), Hyodo, et al. (2015) などを参照）. ここでは非正規母集団の下で, 特に共分散構造が“ ブロック対角構造 ”を持つかどうかの検定について議論し, この問題に対して新たな検定統計量の提案を行った.

いま, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ を平均ベクトル $\boldsymbol{\mu}$, 分散共分散行列 Σ の p 次元母集団からの互いに独立な N 個の観測ベクトルとし, $\boldsymbol{x}_i = \boldsymbol{\mu} + \Gamma \boldsymbol{z}_i$ $(i = 1, \ldots, N)$ を満たすものとする. ここで, Γ は $\Gamma\Gamma' = \Sigma$ を満たす $p \times m$ 行列であり, $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{im})'$ は $\mathrm{E}[\boldsymbol{z}_i] = \boldsymbol{0}$, $\mathrm{Var}[\boldsymbol{z}_i] = I_m$ を満たす互いに独立な m 次元ベクトルである. また, $\boldsymbol{x}_i$, $\boldsymbol{\mu}$, Σ はそれぞれ次のように分割されるものとする.

$$\boldsymbol{x}_i = \begin{pmatrix} \boldsymbol{x}_i^{(1)} \\ \boldsymbol{x}_i^{(2)} \\ \vdots \\ \boldsymbol{x}_i^{(q)} \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \\ \vdots \\ \boldsymbol{\mu}^{(q)} \end{pmatrix}, \ \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1q} \\ \Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{q1} & \Sigma_{q2} & \cdots & \Sigma_{qq} \end{pmatrix}.$$

ここで, $g, h = 1, \ldots, q$ に対して $\boldsymbol{x}^{(g)}$, $\boldsymbol{\mu}^{(g)}$ は p_g 次元ベクトル, Σ_{gh} は $p_g \times p_h$ 行列とし, $\boldsymbol{x}_i^{(g)} = \boldsymbol{\mu}^{(g)} + \Gamma^{(g)} \boldsymbol{z}_i$ $(i = 1, \ldots, N)$ を満たすものとする. ただし, $\Gamma^{(g)}$ は $\Gamma = (\Gamma^{(1)'}, \Gamma^{(2)'}, \ldots, \Gamma^{(q)'})'$ となる $p_g \times m$ 行列である. さらに標本平均ベクトル, 標本共分散行列をそれぞれ

$$\overline{\boldsymbol{x}} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i, \quad S = \frac{1}{N-1} \sum_{i=1}^{N} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})'$$

とし, $\boldsymbol{x}_i^{(g)}$ に対する標本平均ベクトル, 標本共分散行列をそれぞれ

$$\overline{\boldsymbol{x}}^{(g)} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i^{(g)}, \quad S_g = \frac{1}{N-1} \sum_{i=1}^{N} (\boldsymbol{x}_i^{(g)} - \overline{\boldsymbol{x}}^{(g)})(\boldsymbol{x}_i^{(g)} - \overline{\boldsymbol{x}}^{(g)})'$$

とする.

このとき, $\Sigma_d = \mathrm{diag}(\Sigma_{11}, \Sigma_{22}, \cdots, \Sigma_{qq})$ とし, 次の仮説検定問題を考える.

$$H_0 : \Sigma = \Sigma_d \quad \text{vs.} \quad H_1 : \ \Sigma \neq \Sigma_d. \tag{1}$$

一般に, $p < N$ の場合, 共分散構造の検定問題に対して尤度比検定統計量が用いられるが, $p \geq N$ の場合は用いることが出来ない. そのため Hyodo et al. (2015) では, (1) の仮説検定問題に対して, $p \geq N$ の場合でも用いることができる検定方式を正規母集団の下で提案している.

ここでは正規性を仮定しない場合の議論を行うが, 帰無仮説 H_0 の下での共分散行列 Σ_d と真の共分散行列 Σ の間の距離を次の尺度によって測るものとする.

$$\mathrm{tr}(\Sigma - \Sigma_d)^2 = \mathrm{tr}\Sigma^2 - \mathrm{tr}\Sigma_d^2. \tag{2}$$

この (2) の推定量を, Himeno and Yamada (2014) やなどで提案されている $\mathrm{tr}\Sigma^2$ の推定量を用いることによって, 次のように与える.

$$T = \widehat{\mathrm{tr}\Sigma^2} - \widehat{\mathrm{tr}\Sigma_d^2}.$$

ここで,

$$\begin{aligned}
\widehat{\mathrm{tr}\Sigma^2} &= \frac{N-1}{N(N-2)(N-3)}\left\{(N-1)(N-2)\mathrm{tr}S^2 + (\mathrm{tr}S)^2 - NQ\right\} \\
\widehat{\mathrm{tr}\Sigma_d^2} &= \frac{N-1}{N(N-2)(N-3)}\sum_{g=1}^{q}\left\{(N-1)(N-2)\mathrm{tr}S_g^2 + (\mathrm{tr}S_g)^2 - NQ_g\right\}
\end{aligned}$$

であり,

$$\begin{aligned}
Q &= \frac{1}{N-1}\sum_{i=1}^{N}\left\{(\boldsymbol{x}_i - \overline{\boldsymbol{x}})'(\boldsymbol{x}_i - \overline{\boldsymbol{x}})\right\}^2 \\
Q_g &= \frac{1}{N-1}\sum_{i=1}^{N}\left\{(\boldsymbol{x}_i^{(g)} - \overline{\boldsymbol{x}}^{(g)})'(\boldsymbol{x}_i^{(g)} - \overline{\boldsymbol{x}}^{(g)})\right\}^2.
\end{aligned}$$

本報告では, この推定量 T を検定統計量として (1) に対する検定方式を構築するために, 次元と標本サイズが共に大である場合の T の極限分布を, いくつかの高次元枠組みの下で導出した. また, 提案手法の漸近検出力について議論を行った. さらに, 得られた理論結果に対して, モンテカルロ・シミュレーションによって数値的に近似精度や検出力を評価した.

参考文献

[1] Chen, S. X., Zhang, L. X. and Zhong, P. S. (2010). "Tests for high-dimensional covariance matrices", *Journal of the American Statistical Association*, **105,** 810–819.

[2] Himeno, T. and Yamada, T. (2014). "Estimations for some functions of covariance matrix in high dimension under non-normality and its applications", *Journal of Multivariate Analysis*, **130,** 27–44.

[3] Hyodo, M., Shutoh, N., Nishiyama, T. and Pavlenko, T. (2015). "Testing block-diagonal covariance structure for high-dimensional data", *to appear in Statistica Neerlandica.*

[4] Schott, J. R. (2005). "Testing for complete independence in high dimensions", *Biometrika*, **92**, 951–956.

Dirichlet Processを用いた遺伝子発現データのクラスタリング

（公財）がん研究会・ゲノムセンター　牛嶋　大

Email: masaru.ushijima@jfcr.or.jp

1. はじめに

1990 年代初頭，半導体製造技術の応用によりマイクロアレイが開発されて以来，生体標本から数万に及ぶ転写産物の発現情報を取得することが可能となり，網羅的な遺伝子発現解析による生命現象の解明に大きな期待が寄せられた．マイクロアレイデータ解析の主な目的には，(i) 新たな表現型の分類や発見，(ii) 薬剤感受性などの表現型と関連する遺伝子の抽出，(iii) 遺伝子発現データによる表現型の予測モデル構築などがあげられる．論文発表に使われたマイクロアレイデータの多くは公開され，米国 NCBI の GEO (Gene Expression Omnibus) や欧州 EBI の ArrayExpress といった公的データベースに収められていて，誰でも利用することができる．そのため，蓄積されたデータを用いたメタアナリシスなどの研究も盛んに行われている．

2. 遺伝子発現データのクラスタリング

マイクロアレイを用いて得られる網羅的遺伝子発現情報に基づくがんの分子レベルでの分類が可能となった．Sorlie et al.[3] はクラスター解析により，乳がんを 5-6 種のサブタイプに分類し，それらが臨床的転機にも特徴を持つ分類であることを示した．このようにサブタイプに分類することで薬剤感受性や患者の予後に関連する情報が得られれば，個別化医療へとつなげていくことができる．

階層型クラスタリング

遺伝子発現データの教師なし学習として最もよく行われるのがクラスタリングである．階層クラスタリングは，教師なし学習による分類の最も単純でかつよく用いられる方法であり，再帰的に最も「近く」にあるサンプルをグルーピングすることにより実行される．サンプルの「近さ」を定義する距離には Euclid 距離か相関係数を用いるのが一般的である．階層クラスタリングではクラスタ間のグループ分けの順序と類似度を示すデンドログラム（樹形図）が作られるのが特徴であり，マイクロアレイ解析では検体間と遺伝子間で階層クラスタリングを行った後遺伝子の発現量を色で表示したヒートマップがよく作成される．

非階層型クラスタリング

一方，非階層的なクラスタリング手法には K-means 法などがあり，事前に指定したクラスタ数に対し，(i) クラスタの重心の初期値を与える，(ii) 各個体を最も近いクラスタに割り当てる，(iii) 割り当てられた個体からクラスタの重心を再計算する，(ii)-(iii) を収束するまで繰り返すことでクラスタリングが行われる．

一般に事前に与えられたクラスタの数が正しいときは，K-means 法は階層クラスタリングよりよい結果をもたらすことが多い．他方，事前にクラスタの数がわからない場合は，階層クラスタリングが適しているといえる．また，K-means 法の問題点として，クラスタリング結果が重心の初期値に依存する点がある．これに対しては，リサンプリングをしてクラスタリングをくりかえすことで安定的な結果を得るコンセンサスクラスタリング [2] という方法もよく使われる．

クラスター数の推定

クラスタリングに関する問題の一つとして，クラスター数の推定がある．階層型クラスタリングの場合にはその結果からデンドログラムを見てもおおよその判断ができ，また Gap 統計量といったものも提案されている [**?**]．K-means クラスタリングでは先に K の値を与えなければいけないが，コンセンサスクラスタリングでは，K の値を変動させて解析を行い，それぞれの Gini 係数の値から最適なクラスタ数を決定することができる．また，マイクロアレイデータに正規混合モデルを仮定して推定を行い，BIC によって最適なクラスター数を決定する方法も提案されている [1].

3. Dirichlet Processとクラスタリング

Dirichlet Process (DP) を式で書き表すと

$$G \sim \mathrm{DP}(\cdot|G_0, \alpha) \tag{1}$$

のように書け，α はスケーリングパラメータ，G_0 はベースの分布である．G は無限次元で，DP は無限次元のディリクレ分布と考えることができる．

Dirichlet process の応用として最もよくあるのは，混合モデル (mixture model) を用いたデータのクラスタリングである．観測値 $(x_i, \ldots, x_n)$ を潜在パラメータ $(\theta_1, \ldots, \theta_n)$ を用いてモデル化することを考える．個々の θ_i が iid で Dirichlet process G から得られたものとし，また x_i は θ_i をパラメータとする分布 $p(\cdot|\theta_i)$ に従うとすると，

$$\begin{aligned} x_i|\theta_i &\sim p(\cdot|\theta_i) \\ \theta_i|G &\sim G \\ G &\sim \mathrm{DP}(\cdot|G_0, \alpha) \end{aligned} \tag{2}$$

のように表すことができる．G は離散であることから，複数の θ_i が同じ値をとる場合があり，したがって (2) 式は混合モデルと考えることができ，また同じ θ_i に対する x_i は同じクラスターに属すると考えることができる．

Dirichlet Process を用いたクラスタリングの最大の特徴は，推定すると同時にクラスター数 K が決まることである．遺伝子発現データのクラスタリングに適用した論文はいくつか報告されているが [4, 5]，数百ものサンプル数を扱ったものはなく，遺伝子数をかなり減らさないと PC で解析することはできなかった．

4. まとめと今後の課題

我々の目的とするところは，数百から数千のサンプルに対してクラスタリングを行い，得られたサブタイプに特徴的な遺伝子群を抽出することである．今回クラスター数を自動で決めることができる Dirichlet Process によるクラスタリングを試したが，推定アルゴリズムも含めて大量データに対応する工夫が必要であることが明らかになった．今後は Lasso による遺伝子選択も組み入れることを検討し，効率的に解析できるよう改良したい．

参考文献

[1] Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.*, **97**, 611-31.

[2] Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning.* **52**, 91-118.

[3] Sorlie, T., Perou, C. M. Tibshirani, R. et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. USA*, **98**, 10869-74.

[4] Vavoulis, D. V., Francesctto, M., Heutink, P. and Gough, J. (2015). DGEclust: differential expression analysis of clustered count data. *Genome Biol.*, 16:39.

[5] Wang, L. and Wang, X. (2013). Hierarchical Dirichlet process model for gene expression clustering. *EURASIP J. Bioinform. System Biol.*, **2013**:5.

Modeling Circular Markov Processes with Time Varying Autocorrelation

Toshihiro Abe (Nanzan University)
Hiroaki Ogata (Tokyo Metropolitan University)
Takayuki Shiohama (Tokyo Univerisity of Science)
Hiroyuki Taniai (Waseda Univerisity)

1. Introduction

Circular or directional data refers to data recorded as points which directions are measured and arises in biology, geography, medicine, astronomy, and many other areas. Circular data are usually expressed in terms of compass angles or pairs of sine and cosine variables and has the property that the beginning and end of the scale in the domain coincide. By the nature of such periodicity, analyzing circular data is challenging because usual statistics will not be meaningful and will be misleading when applied to circular data without taking into account the particular definition of the domain. Despite the fact that most of the circular data are in the form of time series, not much research has been done in the field of circular time series analysis, and there is still a lot of approaches for development in circular time series modeling.

In general, there are mainly two approaches to model the circular time series; one is the method used to obtain circular-valued random variables by wrapping and the other is Markov models for directional time series. The former approach includes autoregressive circular models (CAR) and linked ARMA model (LARMA) of Fisher and Lee (1994) and the latter includes circular Markov process of Wehrly and Johnson (1979), the Markov process with the Möbius circle transformation of Kato (2010), and Hidden Markov Models (HMM) of Holzmann et al (2006). Recently, Abe et al. (2015) studied the circular Markov process of Wehrly and Johnson (1979), and obtained the theoretical circular autocorrelation structures under the simple model assumptions. According to their results, the circular autocorrelations are completely determined by in terms of the mean resultant length of the underlying circular density of the process.

In this paper, we deal with the circular process of Wehrly and Johnson (1979) which allows for time varying parameters. The proposed models have resulted in allowing us to incorporate time varying autocorrelations of the observed circular time series. Time varying parameters in linear regression and time series models become more and more relevant as the length of the observed time series increases as, and the series itself is subject to changes in the dynamic structure.

The proposed models are used to illustrate how the wind direction and speed are related via the time varying parameters. The time series analysis of wind directions are considered in, for example, Breckling (1989), Ailliot et al. (2006), and Fuentes et al. (2005).

2. Circular Markov Processes

The model considered in this paper is based on the model proposed by Wehrly and Johnson (1979). They proposed a Markov process such that the conditional distribution of θ_t given

by θ_{t-1} is as follows

$$p(\theta_0) = f(\theta_0; \boldsymbol{\omega}_f), \tag{2.1}$$
$$p(\theta_t|\theta_0, \ldots \theta_{t-1}) = p(\theta_t|\theta_{t-1}) = 2\pi g[2\pi\{F(\theta_t; \boldsymbol{\omega}_f) - F(\theta_{t-1}; \boldsymbol{\omega}_f)\}; \boldsymbol{\omega}_g] f(\theta_t; \boldsymbol{\omega}_f) \tag{2.2}$$

where $\boldsymbol{\omega}_f \in \boldsymbol{\Theta}_{\omega_f} \subset \mathbb{R}^{d_f}$ and $\omega_g \in \boldsymbol{\Theta}_{\omega_g} \subset \mathbb{R}^{d_g}$ are the parameters in the circular density $f(\cdot)$ and $g(\cdot)$, respectively. Denote $\boldsymbol{\omega}_g = (\omega_g^{(s)}, \boldsymbol{\omega}_{2,g}^\top)^\top$ where $\omega_g^{(s)} \in \Theta_g \subset \mathbb{R}$ is the scalar scale parameter in $g(\cdot)$ and $\boldsymbol{\omega}_{2,g} \in \boldsymbol{\Theta}_{\omega_{2,g}} \subset \mathbb{R}^{d_g - 1}$ are the nuisance parameters. Let us consider the case that the scale parameter $\omega_g^{(s)}$ of g could be time varying, such that

$$p(\theta_t|\theta_{t-1}) = 2\pi g[2\pi\{F(\theta_t; \boldsymbol{\omega}_f) - F(\theta_{t-1}; \boldsymbol{\omega}_f)\}; (\omega_{g,t}^{(s)}, \boldsymbol{\omega}_{2,g}^\top)] f(\theta_t, \boldsymbol{\omega}_f), \tag{2.3}$$

where $\omega_{g,t}^{(s)}$ is the time dependent scale parameter in g. The scale or concentration parameter $\omega_{g,t}^{(s)}$ is completely determined by an exogenous variable V_t in the following way. Writing $\omega_{g,t}^{(s)}$ as $\omega_{g,t}$, for simplicity. The time varying parameter $\omega_{g,t}^{(s)}$ should be expressed as follows

$$\omega_{g,t} = h(m(V_t))$$

where $h : \mathbb{R} \to \boldsymbol{\Theta}_{\omega_g}$ is the known function and chosen such that the $\omega_{g,t}^{(s)}$ satisfies the parameter restrictions of $\boldsymbol{\Theta}_{\omega_g}$, and $m : \mathbb{R}^+ \to \mathbb{R}$ is the unknown function. The goal in this study is to estimate unknown parameter vectors $(\boldsymbol{\omega}_f^\top, \boldsymbol{\omega}_{2,g}^\top)^\top$ and the function $m(\cdot)$.

In this study, we discuss about the maximum likelihood estimation for the proposed models and its asymptotics. The proposed models are illustrated using wind direction and speed in Wakkanai, Hokkaido, Japan.

References

Abe, T. and Pewsey, A. (2011). Sine-skewed circular distributions. *Statistical Papers* **52**, 683–707.

Ailliot, P., Monbet, V., and Prevosto, M. (2006). An autoregressive model with time-varying coefficients for wind fields. *Environmetrics*, **17**, 107–117.

Abe, T., Ogata, H., Shiohama, T., and Taniai, H. (2015). A circular autocorrelation of stationary circular Markov processes. Working Paper.

Breckling, J. (1989). *The analysis of directional time series: application to wind speed and direction.* Lecture Notes in Statistics, **61**, Springer-Verlag, Berlin.

Fisher, N. I. and Lee, A. J. (1994). Regression models for an angular response. *Biometrics*, **48**, 665–677.

Fisher, N. I. and Lee, A. J. (1994). Time series analysis of circular data. *Journal of the Royal Statistic Society, Series B*, **70**, 327–332.

Fuentes, M., Chen, L., Davis, J. M., and Lachmann, G. M. (2005). Modeling and predicting complex space-time structures and patterns of coastal wind fields. *Environmetrics*, **16**, 449–464.

Holzmann, H., Munk, A., Suster, M., and Zucchini, W. (2006). Hidden Markov models for circular and linear-circular time series. *Environmental Ecological Statistics*, **13**, 325–347.

Hardle, W., Hall, P., and Ichimura, H. (1993). Optimal smoothing in single index models. *Annals of Statistics*, **21**, 157–178.

Kato, S. (2010). A markov process for circular data. *Journal of the Royal Statistical Society: Series B*, **72**, 655–672.

Wehrly, T. E. and Johnson, R. A. (1979). Bivariate models for dependence of angular observations and a related Markov process. *Biometrika*, **66**, 255–256.

Statistical inference for nonstationary cluster point processes: an extension of Tanaka-Ogata's Palm likelihood method

非定常空間クラスター点過程におけるパラメータ推定：田中・尾形の **Palm** 尤度法の拡張

Shimatani, Ichiro K. (shimatan@ism.ac.jp): The Institute of Statistical Mathematics, 10-3, Midori, Tachikawa, Tokyo 190-8562.
島谷健一郎 (統計数理研究所)

Introduction

A spatial distribution of forest trees is a typical example of spatial point patterns. In general, tree densities changes along some environmental gradients, and trees often exhibit clustering due to limited seed dispersal from a mother tree.
Such point patterns can be modeled by the inhomogeneous Neyman-Scott process. For this spatial point process model, simulating an artificial dataset is easy whereas optimizing parameter values is difficult. In order to estimate parameter values from given data, this study extended the method proposed by Tanaka et al. (2008), called as the maximum log Palm likelihood method, and showed its performance.

Inhomogeneous Neyman-Scott process

Suppose that (1) a parental population followed the inhomogeneous Poisson process of intensity $d(\mathbf{x}; \theta)$; (2) each parent produced u daughters where u is a random number from the Poisson distribution of intensity ν; (3) daughters were dispersed from each mother tree according to the two-dimensional Gaussian distribution $f(r; \sigma^2) = \exp(-r^2/2\sigma^2)/2\pi\sigma^2$; (4) there was no interaction among daughters; (5) parents all died; and (6) the survival probability of a daughter at $\mathbf{x}$ is given by $s(\mathbf{x}; \varphi)$.

The resulting point pattern is an example of inhomogeneous Neyman-Scott processes.

The 1st-order intensity is given by

$$\rho(x;\theta,\nu,\sigma^2,\varphi) = \int d(z;\theta)E(u)\, f(\| x-z \|;\sigma^2)dz \cdot s(x;\varphi)\,.$$

The 2nd-order moment function is given by

$$\rho^{(2)}(\mathbf{x},\mathbf{y}) = \rho(x;\theta,\nu,\sigma^2,\varphi)\rho(y;\theta,\nu,\sigma^2,\varphi) + \int d(x;\theta)E(u^2)\, f(\| \mathbf{x}-\mathbf{z} \|;\sigma^2)\, f(\| \mathbf{y}-\mathbf{z} \|;\sigma^2)d\mathbf{z}$$

Extending the homogeneous case, we introduce the local Palm intensity function at $\mathbf{x}$ as the occurrence probability of another daughter at $\mathbf{y}$ given that the presence at $\mathbf{x}$:

$$\Lambda_x(y) = \rho^{(2)}(x,y) / \rho(x)\,.$$

If we approximate the point distribution on the square centered at **x** and edge length of *R* by the inhomogeneous Poisson process with intensity function $\Lambda_x(y)$, the log-likelihood is given by:

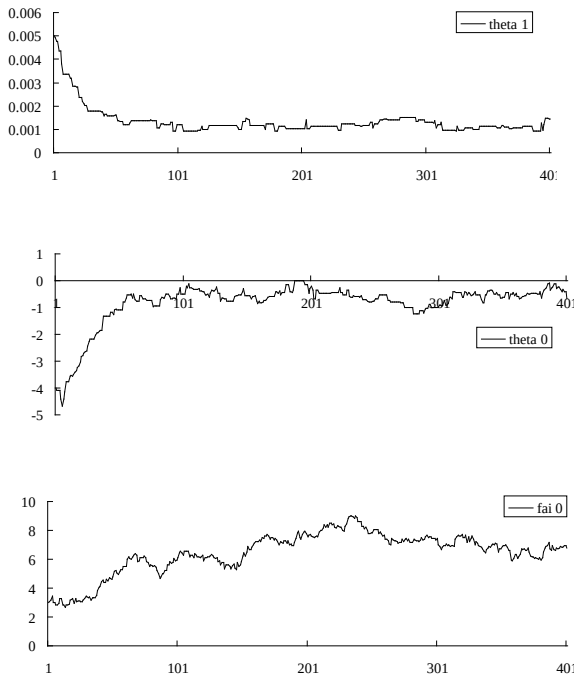
$$\sum_{\|y_j-\mathbf{x}\|_M \le R} \ln(\Lambda_x(y_j)) - \int_{\|\mathbf{y}-x\|_M \le R} \Lambda_x(\mathbf{y}) d\mathbf{y}$$

in which, $\| \|_M$ indicates the Manhattan metric.

Modifying the log Palm likelihood functions, on the assumption of the independence of the above (approximated) Poisson processes over $\mathbf{x}_i \in B$, we introduced the extended log Palm likelihood by:

$$\ln \tilde{L}(\mu,\nu,\sigma,\Theta) = \sum_{i\in B} \{ \sum_{\|\mathbf{x}_j-\mathbf{x}_i\|_M \le R} \ln(\Lambda_{\mathbf{x}_i}(\mathbf{x}_j)) - \int_{\|\mathbf{y}-\mathbf{x}_i\|_M \le R} \Lambda_{\mathbf{x}_i}(\mathbf{y}) d\mathbf{y} \} .$$

Below shows examples of parameter estimation by the Metropolis-Hasting algorithm. Basically, the method is working efficiently.



Literatures cited

Tanaka, U., Ogata, Y. & Stoyan, D. 2008. Parameter estimation and model selection for Neyman-Scott point processes. *Biometrical Journal*, 50, 43–57.

順序制約がある２つの正規母平均の推定－分散共分散行列が既知の場合

目白大学　張　元宗　　慶應大学　篠崎 信雄

主旨：　２次元正規分布の分散共分散行列が既知で、順序制約がある母平均の推定問題を考える。この問題に対して、Hwang and Peddada(1994) または、Peddada et al. (2005) が提案した推定量の妥当性はあまり明らかにされていないため、ここでは、確率優越性の評価基準の下で、制約条件を満たす最尤推定量が Hwang and Peddada(1994)、または Peddada et al. (2005) が提案した推定量より優れていることを明らかにする。また、Pitman nearness の評価基準の下でも同様な結果が得られる。

1. はじめに：　$\boldsymbol{X} = (X_1, X_2)' \sim N(\boldsymbol{\mu}, \Sigma)$ に従い、分散共分散行列

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}, \tag{1.1}$$

は既知で、$|\rho| \neq 1$、$\boldsymbol{\mu} = (\mu_1, \mu_2)'$ とし、$\mu_i, i = 1, 2$ に順序制約, $\mu_1 \leq \mu_2$, がある場合 $\mu_i, i = 1, 2$ の推定を考える。$\rho = 0$ の場合、μ_i の最尤推定量は

$$\hat{\mu}_1 = \min\left\{\bar{X}_1, \frac{\sigma_2^2\bar{X}_1 + \sigma_1^2\bar{X}_2}{\sigma_1^2 + \sigma_2^2}\right\},\ \hat{\mu}_2 = \max\left\{\bar{X}_2, \frac{\sigma_2^2\bar{X}_1 + \sigma_1^2\bar{X}_2}{\sigma_1^2 + \sigma_2^2}\right\}. \tag{1.2}$$

である。$\rho \neq 0$ の場合、Hwang, Peddada (1994) は $\hat{\mu}_i$ を拡張した $\boldsymbol{\mu}$ の推定量

$$\hat{\mu}_1^{HP} = \min\left\{\bar{X}_1, \alpha\bar{X}_1 + \beta\bar{X}_2\right\},\ \hat{\mu}_2^{HP} = \max\left\{\bar{X}_2, \alpha\bar{X}_1 + \beta\bar{X}_2\right\}. \tag{1.3}$$

を提案、$\hat{\mu}_i^{HP}$ は確率的に $\bar{X}_i, i = 1, 2$ より優れていることを証明した。ここで、$\alpha = \omega_1/(\omega_1 + \omega_2)$, $\beta = \omega_2/(\omega_1 + \omega_2)$、$\omega_1 = \sigma_2^2 - \rho\sigma_1\sigma_2$, $\omega_2 = \sigma_1^2 - \rho\sigma_1\sigma_2$, $|\rho| \neq 1$ であるので、$\omega_1 + \omega_2 = (\sigma_1 - \sigma_2)^2 + 2(1 - \rho)\sigma_1\sigma_2 > 0$. である。Peddada et al. (2005) は $\omega_1\omega_2 < 0$ の場合、推定量 $\hat{\mu}_i^{HP}, i = 1, 2$ は一致推定量にならないことに気づき、$\hat{\mu}_i^{HP}$ を次のように修正し、確率的に $\bar{X}_i$ より優れていることを証明した。

$$\hat{\mu}_1^{PDT} = \min\left\{\bar{X}_1, \alpha^*\bar{X}_1 + \beta^*\bar{X}_2\right\},\ \hat{\mu}_2^{PDT} = \max\left\{\bar{X}_2, \alpha^*\bar{X}_1 + \beta^*\bar{X}_2\right\}, \tag{1.4}$$

ここで、$\alpha^* = \omega_1^+/(\omega_1^+ + \omega_2^+)$, $\beta^* = \omega_2^+/(\omega_1^+ + \omega_2^+)$, $a^+ = \max\{a, 0\}$. しかし、提案した両推定量の妥当性はあまり調べられておらず、不自然な場合がある。たとえば、$\bar{X}_1 > \bar{X}_2$ および $\omega_1 + \omega_2 > 0, \omega_2 < 0$ の場合、制約条件を満たしていないのに、μ_1 の推定量を $\hat{\mu}_1^{HP} = \bar{X}_1$ にするような不自然な推定量。逆に、$\bar{X}_2 > \bar{X}_1$ の場合、制約条件を満たしているのに、$\bar{X}_1$ を原点に縮小するような不自然な推定量 $\hat{\mu}_1^{HP} = \bar{X}_1 + \beta(\bar{X}_2 - \bar{X}_1)$。同様に、このようなことは $\hat{\mu}_1^{PDT}$ に対しても起こる。ここでは、制約条件を満たす $\boldsymbol{\mu}$ の最尤推定量

$$\hat{\mu}_1^{MLE} = \bar{X}_1 - \beta(\bar{X}_1 - \bar{X}_2)^+,\ \hat{\mu}_2^{MLE} = \bar{X}_2 + \alpha(\bar{X}_1 - \bar{X}_2)^+, \tag{1.5}$$

と両推定量 $\hat{\mu}_i^{HP}, \hat{\mu}_i^{PDT}$ との比較を行い、両推定量より確率的に優れていることを明らかにする。
まず、$\boldsymbol{y} = (y_1, y_2)'$ とし、2 次元平面の 3 つの領域を

$$D_1 = \{\boldsymbol{y} | y_1 + y_2 > 0, y_1 \geq 0, y_2 \geq 0\},\ D_2 = \{\boldsymbol{y} | y_1 + y_2 > 0, y_2 < 0\},\ D_3 = \{\boldsymbol{y} | y_1 + y_2 > 0, y_1 < 0\}.$$

とすると $\boldsymbol{\omega} = (\omega_1, \omega_2)'$ は D_1, D_2, D_3 にいずれか属し、3 推定量の関係はつぎのように整理される。

$$\begin{aligned}
\boldsymbol{\omega} \in D_1 &: \hat{\mu}_1^{MLE} = \hat{\mu}_1^{HP} = \hat{\mu}_1^{PDT} = \bar{X}_1 - \beta(\bar{X}_1 - \bar{X}_2)^+, \\
&\quad\ \hat{\mu}_2^{MLE} = \hat{\mu}_2^{HP} = \hat{\mu}_2^{PDT} = \bar{X}_2 + \alpha(\bar{X}_1 - \bar{X}_2)^+, \\
\boldsymbol{\omega} \in D_2 &: \hat{\mu}_1^{MLE} = \bar{X}_1 - \beta(\bar{X}_1 - \bar{X}_2)^+, \hat{\mu}_1^{HP} = \bar{X}_1 + \beta(\bar{X}_2 - \bar{X}_1)^+, \hat{\mu}_1^{PDT} = \bar{X}_1; \\
&\quad\ \hat{\mu}_2^{MLE} = \hat{\mu}_2^{HP} = \bar{X}_2 + \alpha(\bar{X}_1 - \bar{X}_2)^+, \hat{\mu}_2^{PDT} = \bar{X}_2 + (\bar{X}_1 - \bar{X}_2)^+, \\
\boldsymbol{\omega} \in D_3 &: \hat{\mu}_1^{MLE} = \hat{\mu}_1^{HP} = \bar{X}_1 - \beta(\bar{X}_1 - \bar{X}_2)^+, \hat{\mu}_1^{PDT} = \bar{X}_1 - (\bar{X}_1 - \bar{X}_2)^+; \\
&\quad\ \hat{\mu}_2^{MLE} = \bar{X}_2 + \alpha(\bar{X}_1 - \bar{X}_2)^+, \hat{\mu}_2^{HP} = \bar{X}_2 - \alpha(\bar{X}_2 - \bar{X}_1)^+, \hat{\mu}_2^{PDT} = \bar{X}_2.
\end{aligned}$$

２．結果：　**定理１．**　$\hat{\mu}_i^{MLE}$ は確率的に $\hat{\mu}_i^{HP}, i = 1, 2,$ より優れている。つまり、すべての $d > 0,$ に対して、

$$P_r\{|\hat{\mu}_i^{MLE} - \mu_i| < d\} \geq P_r\{|\hat{\mu}_i^{HP} - \mu_i| < d\}$$

が成立する。

証明：まず、$\hat{\mu}_1^{MLE}$ は $\hat{\mu}_1^{HP}$ より確率的に優れていること証明する。$\boldsymbol{\omega} \in D_2$ のみで $\hat{\mu}_1^{MLE}$ と $\hat{\mu}_1^{HP}$ が異なるので、$\boldsymbol{\omega} \in D_2$ で、すべての $d > 0$ に対して

$$P\{|\hat{\mu}_1^{MLE} - \mu_1| < d\} \geq P\{|\hat{\mu}_1^{HP} - \mu_1| < d\}$$

を証明すればよい。$P\{|\hat{\mu}_1^{MLE} - \mu_1| < d\}$ を次のように評価する。

$$\begin{aligned}
&P\{|\hat{\mu}_1^{MLE} - \mu_1| < d\} \\
&= P\{|\bar{X}_1 - \mu_1| < d, \bar{X}_2 \geq \bar{X}_1\} + P\{|\alpha\bar{X}_1 + \beta\bar{X}_2 - \mu_1| < d, \bar{X}_2 < \bar{X}_1\} \\
&= P\{-d < \bar{X}_1 - \mu_1 < d, \bar{X}_2 \geq \bar{X}_1\} + P\{-d < \alpha(\bar{X}_1 - \mu_1) + \beta(\bar{X}_2 - \mu_1) < d, \bar{X}_2 < \bar{X}_1\}
\end{aligned} \tag{2.1}$$

次のような変数変換

$$V_i = \bar{X}_i - \mu_1,\ i = 1, 2, \tag{2.2}$$

を行うと

$$\begin{pmatrix} V_1 \\ V_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ \Delta \end{pmatrix}, \begin{bmatrix} \frac{\sigma_1^2}{n} & \frac{\rho\sigma_1\sigma_2}{n} \\ \frac{\rho\sigma_1\sigma_2}{n} & \frac{\sigma_2^2}{n} \end{bmatrix} \right) \tag{2.3}$$

に従う。ここで $\Delta = \mu_2 - \mu_1 \geq 0.$ である。よって、(2.1) は

$$P\{|\hat{\mu}_1^{MLE} - \mu_1| < d\} = P\{-d < V_1 < d, V_2 \geq V_1\} + P\{-d < \alpha V_1 + \beta V_2 < d, V_2 < V_1\}. \tag{2.4}$$

になる。さらに、変数変換

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} \alpha & \beta \\ 1 & -1 \end{pmatrix} \begin{pmatrix} V_1 \\ V_2 \end{pmatrix} = \begin{pmatrix} \alpha V_1 + \beta V_2 \\ V_1 - V_2 \end{pmatrix} \tag{2.5}$$

を行うと $Z_1 \sim N(\beta\Delta, \sigma^2),\ Z_2 \sim N(-\Delta, (\omega_1 + \omega_2)/n)$ に従い、$cov(Z_1, Z_2) = 0$ になる。ここで、$\sigma^2 = (\alpha^2\sigma_1^2 + 2\alpha\beta\rho\sigma_1\sigma_2 + \beta^2\sigma_2^2)/n.$ である。よって、$V_1 = Z_1 + \beta Z_2,\ V_2 = Z_1 - \alpha Z_2$ になり、(2.4) は

$$\begin{aligned}
&P\{-d < V_1 < d, V_2 \geq V_1\} + P\{-d < \alpha V_1 + \beta V_2 < d, V_2 < V_1\} \\
&= P\{-d < Z_1 + \beta Z_2 < d, Z_2 \leq 0\} + P\{-d < Z_1 < d, Z_2 > 0\} \\
&= P\left\{ \frac{-d - \beta(Z_2 + \Delta)}{\sigma} < \frac{Z_1 - \beta\Delta}{\sigma} < \frac{d - \beta(Z_2 + \Delta)}{\sigma}, Z_2 \leq 0 \right\} + P\left\{ \frac{-d - \beta\Delta}{\sigma} < \frac{Z_1 - \beta\Delta}{\sigma} < \frac{d - \beta\Delta}{\sigma}, Z_2 > 0 \right\} \\
&= \int_{-\infty}^{0} g_\beta(z_2 + \Delta) f(z_2) dz_2 + g_\beta(\Delta) \int_0^{\infty} f(z_2) dz_2
\end{aligned}$$

になる。ここで

$$g_\beta(z) = \Phi\left(\frac{d - \beta z}{\sigma} \right) - \Phi\left(\frac{-d - \beta z}{\sigma} \right)$$

であり、$f(z_2)$ は z_2 の密度関数である。同様に、$\boldsymbol{\omega} \in D_2$ で、$P\{|\hat{\mu}_1^{HP} - \mu_1| < d\}$ は

$$\begin{aligned}
P\{|\hat{\mu}_1^{HP} - \mu_1| < d\} &= P\{-d < \alpha\bar{X}_1 + \beta\bar{X}_2 - \mu_1 < d, \bar{X}_2 \geq \bar{X}_1\} + P\{-d < \bar{X}_1 - \mu_1 < d, \bar{X}_2 < \bar{X}_1\} \\
&= g_\beta(\Delta) \int_{-\infty}^{0} f(z_2) dz_2 + \int_0^{\infty} g_\beta(z_2 + \Delta) f(z_2) dz_2
\end{aligned}$$

になる。よって、$\boldsymbol{\omega} \in D_2$ での両推定量の確率の差は

$$\begin{aligned}
\Delta P &= P\{|\hat{\mu}_1^{MLE} - \mu_1| < d\} - P\{|\hat{\mu}_1^{HP} - \mu_1| < d\} \\
&= \int_{-\infty}^{0} \{g_\beta(Z_2 + \Delta) - g_\beta(\Delta)\} f(z_2) dz_2 + \int_0^{\infty} \{g_\beta(\Delta) - g_\beta(Z_2 + \Delta)\} f(z_2) dz_2
\end{aligned} \tag{2.6}$$

になる。Z_2 の分布を原点に変換すると $S = Z_2 + \Delta \sim N(0, (\omega_1 + \omega_2)/n)$ に従い、(2.6) は

$$\begin{aligned}
\Delta P &= \int_{-\infty}^{\Delta} (g_\beta(s) - g_\beta(\Delta)) f(s) ds + \int_{\Delta}^{\infty} (g_\beta(\Delta) - g_\beta(s)) f(s) ds \\
&= \int_{-\infty}^{0} (g_\beta(s) - g_\beta(\Delta)) f(s) ds + \int_0^{\infty} (g_\beta(\Delta) - g_\beta(s)) f(s) ds + 2\int_0^{\Delta} (g_\beta(s) - g_\beta(\Delta)) f(s) ds \\
&= 2\int_0^{\Delta} (g_\beta(s) - g_\beta(\Delta)) f(s) ds \geq 0
\end{aligned}$$

になる。最後の式は下記の理由で成立する。$g_\beta(s)$ は原点に対称で、また、$g_\beta(s)$ は $(-\infty, 0)$ 間は増加関数であり、$(0, \infty)$ の間は減少関数である。

$\hat{\mu}_1^{MLE}$ は μ_1^{HP} より優れていることを証明した。μ_2 の推定に対して、同様に、$\hat{\mu}_2^{MLE}$ は $\hat{\mu}_2^{HP}$ より優れることを証明するため、$\boldsymbol{\omega} \in D_3$ で、両推定量の評価を行えばよい。

次に、 $\hat{\mu}_i^{MLE}$ は確率的に $\hat{\mu}_i^{PDT}, i = 1, 2,$ より優れていることを次の定理にまとめる。

定理 2. $\hat{\mu}_i^{MLE}$ は確率的に $\hat{\mu}_i^{PDT}, i = 1, 2,$ より優れている。つまり、すべての $d > 0,$ に対して、

$$P_r\{|\hat{\mu}_i^{MLE} - \mu_i| < d\} \geq P_r\{|\hat{\mu}_i^{PDT} - \mu_i| < d\}$$

が成立する。

定理 1 と同様に、評価を行えばよい。詳細を省略する。

また、Pitman nearness の評価基準の下でも同様な結果が得られた。

比のカーネル型推定量の漸近的性質について

森山 卓　　(九州大学大学院数理学府)[*1]
前園 宜彦　(九州大学数理学研究院)[*2]

1. 序

統計学において，比の統計量には重要なものが多くある．今回はその中の密度比とハザード比について議論する．密度比はいわゆる尤度比と考えることができ，その応用は幅広いものである．例えば二標本の等分布の検定，変化点の検出，判別分析などがある．互いに独立な分布の密度 $f(x)$，$g(x)$ の点 x_0 での比 $f(x_0)/g(x_0)$ のノンパラメトリックで自然な推定量は $\widehat{f}(x_0)/\widehat{g}(x_0)$ で定義される．ここで $\widehat{f}(x_0)$ は通常のカーネル密度推定量である．カーネルの台を例えば $(-\infty,\infty)$ ととると，分母の $\widehat{g}(x_0)\neq 0$ を保証できる．この推定量の漸近平均二乗誤差 $(AMSE)$ は Chen et al.(2009) によって調べられているが本講演では高次の漸近理論について議論する．

密度比 $f(x_0)/g(x_0)$ の他の推定量としては "direct" estimator が Cwik and Mielniczuk (1989) によって提案されている．我々はこの推定量のアイデアを用いて新たなハザード比の推定量を提案する．ハザード比は $f(x_0)/(1-F(x_0))$ で定義され，この推定は生存時間解析において基本的なものである．ハザード比はいわゆる '死' や破産などのイベントの発生の条件付き確率 $(\lim_{\Delta t\to 0} P(t\leq T<t+\Delta t|T\geq t)/\Delta t)$ を意味する．我々は新たな "direct" hazard 推定量の $AMSE$ について議論を行う．

2. 密度比の推定量

$X_1,\cdots,X_m$ と $Y_1,\cdots,Y_n$ を $i.i.d.$ 確率変数とする．X_i は共通の分布 F，Y_j は G，またその密度を f，g とし $g(x_0)\neq 0$ を仮定する．この密度比の自然なカーネル推定量は $\widehat{f}(x_0)/\widehat{g}(x_0)$ で定義される．ここで $\widehat{f}(x_0)$，$\widehat{g}(x_0)$ は次のように表せる．

$$\widehat{f}(x_0)=\frac{1}{mh_{f,m}}\sum_{i=1}^{m}K_f\left(\frac{x_0-X_i}{h_{f,m}}\right)\qquad \widehat{g}(x_0)=\frac{1}{nh_{g,n}}\sum_{j=1}^{n}K_g\left(\frac{x_0-Y_j}{h_{g,n}}\right)$$

この推定量のエッジワース展開について以下が成り立つ．

Theorem 2.1 いくつかの正則条件の下，$m+n=N$，$O(m)=O(n)$，カーネル K_f,K_g は対称，$h_{f,m}=O(N^{-c}),h_{g,n}=O(N^{-d})$ $(2/13<c<1/2,\ 2/13<d<1/2)$ とすると次のエッジワース展開が有効である．

$$\sup_{-\infty<y<\infty}\left|P\left(V^{-\frac{1}{2}}\left(\frac{\widehat{f}(x_0)}{\widehat{g}(x_0)}-E\right)<y\right)-\Phi(y)\right|=o\left(n^{-\frac{1}{2}}\right),$$

本研究は科研費 (課題番号:15K11995) の助成を受けたものである。
2010 Mathematics Subject Classification: MSC-62G10, MSC-62G20
キーワード：カーネル法，密度比，エッジワース展開，ハザード比，平均二乗誤差
[*1] 〒819-0395 福岡市西区元岡 744　九州大学 大学院数理学府
e-mail: moritaku3542168@gmail.com
[*2] 〒819-0395 福岡市西区元岡 744　九州大学 大学院数理学研究院
e-mail: maesono@math.kyushu-u.ac.jp

ここで,

$$E = \frac{1}{g(x_0)}\left(f(x_0) + h_{f,m}^2 B_{f,2} + h_{f,m}^4 B_{f,4}\right) - \frac{1}{g^2(x_0)}\big(f(x_0)(h_{g,n}^2 B_{g,2} + h_{g,n}^4 B_{g,4}) + h_{f,m}^2 h_{g,n}^2 B_{f,2} B_{g,2}\big) + \frac{1}{nh_{g,n}} \frac{f(x_0)}{g^2(x_0)} \frac{A_{2,0}^g}{2}$$

$$V = V[\widehat{f}(x_0)]\left(\frac{1}{g^2(x_0)} - 2h_{g,n}^2 \frac{1}{g^3(x_0)} B_{g,2}\right) + V[\widehat{g}(x_0)]\left(\frac{f^2(x_0)}{g^4(x_0)} + 2h_{f,m}^2 \frac{f(x_0)}{g^4(x_0)} B_{f,2} - 4h_{g,n}^2 \frac{f^2(x_0)}{g^5(x_0)} B_{g,2}\right),$$

$$A_{i,j}^{\gamma} = \int K_{\gamma}^i(u) u^j du, \quad B_{\gamma,k} = \frac{1}{k!} A_{0,k}^{\gamma} \gamma^{(k)}(x_0)$$

であり， Φ は標準正規分布関数を表すとする.

3. Direct Hazard 推定量

$X_1, \cdots, X_N$ を *i.i.d.* 確率変数とし，その分布を F，密度を f とし $f(x_0) \neq 0$，すなわち $1 - F(x_0) \neq 0$ を仮定する．Cwik and Mielniczuk (1989) のアイデアを利用した新たな 'direct' hazard 推定量を次のように定義する.

$$\widehat{\frac{f}{1-F}}(x_0) = \frac{1}{h_N} \int K\left(\frac{y - \Delta_N(y) - (x_0 - \Delta_N(x_0))}{h_N}\right) dF_N(y)$$

ここで, $F_N(y) = N^{-1} \sum I(X_i \leq y)$ であり, $\Delta_N(y) = \int_{-\infty}^{y} F_N(u) du = N^{-1} \sum I(X_i \leq y)(y - X_i)$ とする.

Theorem 3.1 分布 F とカーネル K に適当な条件をおくと，$h_N = O(N^{-c})$ $(2/13 < c < 1/2)$ の下，$\widehat{\frac{f}{1-F}}(x_0)$ の平均二乗誤差は次のようになる.

$$E\left[\widehat{\frac{f}{1-F}}(x_0) - \frac{f}{1-F}(x_0)\right]^2 = \frac{h_N^4}{4} A_{0,2}^2 \frac{\{(1-F)\{(1-F)f'' + 4ff'\} + 3f^3\}^2}{(1-F)^{10}}(x_0) + \frac{1}{Nh_N} \frac{f}{1-F}(x_0) A_{2,0} + O(h_N^6) + O\left(\frac{1}{Nh_N^{1/2}}\right).$$

参考文献

[1] Chen, S. M., Hsu, Y. S., and Liaw, J. T. (2009). On kernel estimators of density ratio. *Statistics*, **43**, 463-479.

[2] Ćwik, J., and Mielniczuk, J. (1989). Estimating density ratio with application to discriminant analysis. *Communications in Statistics-Theory and Methods*, **18**, 3057-3069.

[3] García-Soidán, P. H., González-Manteiga, W., and Prada-Sánchez, J. (1997). Edgeworth expansions for nonparametric distribution estimation with applications. *Journal of statistical planning and inference*, **65**, 213-231.

[4] Maesono, Y. (1985). Edgeworth expansion for two-sample U-statistics. *Rep. Fac. Sci. Kagoshima Univ*, **18**, 35-43.

[5] Patil, P. N. (1993). On the least squares cross-validation bandwidth in hazard rate estimation. *The Annals of Statistics*, **21**, 1792-1810.

First principal component and its applications to tests of means and covariance matrices for high-dimensional data

筑波大学・数理物質科学研究科　　石井　晶

1. はじめに

情報化の進展に伴い，高次元データの統計的な解析が益々重要になってきている．データの次元数 p が標本数 n よりも遥かに大きな高次元小標本においては，従来の多変量解析の理論は崩壊する．Aoshima and Yata (2011) は，二標本問題の検出力を保証する検定方式や正判別確率を保証する判別方式など 8 つの統計的推測について，高次元データの母集団の差異を幾何学的表現で捉える先駆的な理論と方法論を与えた．

高次元データの特徴として，次元数が大きくなるにつれ，共分散行列の最大固有値がその他の固有値よりも遥かに大きくなることが挙げられる．したがって，第 1 主成分の漸近的な性質をとらえることは，高次元データ解析にとって極めて重要である．本報告では，高次元小標本漸近理論を次元数 $p \to \infty$ だがデータ数 n は固定のもとで展開し，高次元 PCA を用いた新たな統計的推測について述べた．平均に p 次のベクトル　，共分散行列に p 次の非負定値対称行列 $\boldsymbol{\Sigma}$ をもつ母集団を考える．n 個の p 次データベクトル $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ を無作為に抽出して，p　n データ行列 $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n]$ を定義する．ただし，$p > n$ である．$\boldsymbol{\Sigma}$ の固有値を λ_1　　λ_p(　0) とし，$\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, ..., \lambda_p)$ とおく．対応する直交行列 $\boldsymbol{H} = [\boldsymbol{h}_1, ..., \boldsymbol{h}_p]$ で $\boldsymbol{\Sigma} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^T$ と固有値分解する．標本共分散行列を $\boldsymbol{S} = (n \quad 1)^{-1}(\boldsymbol{X} \quad \overline{\boldsymbol{X}})(\boldsymbol{X} \quad \overline{\boldsymbol{X}})^T$ とおく．ここで，$\overline{\boldsymbol{X}} = [\bar{\boldsymbol{x}}, ..., \bar{\boldsymbol{x}}]$，$\bar{\boldsymbol{x}} = \sum_{j=1}^{n} \boldsymbol{x}_j/n$ である．双対な標本共分散行列 $\boldsymbol{S}_D = (n \quad 1)^{-1}(\boldsymbol{X} \quad \overline{\boldsymbol{X}})^T(\boldsymbol{X} \quad \overline{\boldsymbol{X}})$ について，固有値 $\hat{\lambda}_1$　$\hat{\lambda}_{n-1}$(　0) と対応する固有ベクトル $\hat{\boldsymbol{u}}_j,\ j = 1, ..., n \quad 1$ によって，$\boldsymbol{S}_D = \sum_{j=1}^{n-1} \hat{\lambda}_j \hat{\boldsymbol{u}}_j \hat{\boldsymbol{u}}_j^T$ と固有値分解する．

2. 第 1 固有空間の推定

Yata and Aoshima (2012) のノイズ掃き出し法を用いると，最大固有値の推定量は $\tilde{\lambda}_1 = \hat{\lambda}_1 \quad (n \quad 2)^{-1}\{\mathrm{tr}(\boldsymbol{S}_D) \quad \hat{\lambda}_1\}$ で与えられる．Ishii et al. (2015) は，ノイズ掃き出し法を用いて推定した最大固有値に対し，以下の定理を与えた．

定理 1　適当な正則条件のもと，$p \to \infty$，n 固定で以下が成り立つ．

$$(n \quad 1)\frac{\tilde{\lambda}_1}{\lambda_1} \Rightarrow \chi^2_{n-1}$$

ここで，$\Rightarrow$ は分布収束，χ^2_{n-1} は自由度 $n \quad 1$ の χ^2 分布に従う確率変数を表す．

ノイズ掃き出し法による第 1 固有ベクトルの推定量は，$\tilde{h}_1 = \{(n \quad 1)\tilde{\lambda}_1\}^{-1/2}(\boldsymbol{X} \quad \bar{\boldsymbol{X}})\hat{\boldsymbol{u}}_1$ で与えられる．第 1 固有ベクトルについて，Ishii et al. (2015) は以下の定理を与えた．

定理 2　適当な正則条件のもと，$p \to \infty$，n 固定で以下が成り立つ．

$$\boldsymbol{h}_1^T\tilde{\boldsymbol{h}}_1 = 1 + o_p(1).$$

Ishii et al. (2015) では，主成分スコアの推定についても同様に定理を与えている．

3. 平均ベクトルの検定

高次元データに対し，Hotelling の T^2 統計量を用いることはできない．平均ベクトルの

検定 $H_0:\quad =\quad _0$ vs. $H_1:\quad \neq\quad _0$ について，検定統計量を次のように定義する．

$$F_0 = \frac{n||\bar{\boldsymbol{x}}\qquad _0||^2\quad \mathrm{tr}(\boldsymbol{S}_D)}{\tilde{\lambda}_1} + 1$$

定理 3　適当な正則条件のもと，$p \to \infty$，n 固定で以下が成り立つ．

$$F_0 \Rightarrow F_{1,n-1} \text{ under } H_0$$

ここで，F_{ν_1,ν_2} は自由度 (ν_1, ν_2) の F 分布に従う確率変数を表す．

定理 3 に基づいて，$\alpha \in (0, 1/2)$ に対して，検定方式を次のように与える．

$$F_0 \quad F_{1,n-1}(\alpha) \text{ ならば } H_0 \text{ を棄却}$$

ここで，$F_{\nu_1,\nu_2}(\alpha)$ は自由度 (ν_1, ν_2) の F 分布における上側 α 点を表す．

4. 共分散行列の同等性検定

母集団が 2 つある場合を考える．各母集団の共分散行列を $\boldsymbol{\Sigma}_1,\ \boldsymbol{\Sigma}_2$ とおき，次の検定を考える．

$$H_0:\ \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 \quad \text{vs.} \quad H_1:\ \boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2.$$

各母集団から大きさ $n_1,\ n_2$ の無作為標本を抽出する．各母集団について，ノイズ掃き出し法を用い，最大固有値と固有ベクトルの推定量 $\tilde{\lambda}_{1(i)}, \tilde{\boldsymbol{h}}_{1(i)}\ (i = 1, 2)$ を得る．さらに，ノイズ空間 $\kappa_i = \mathrm{tr}(\boldsymbol{\Sigma})\quad \lambda_1$ の推定量を $\tilde{\kappa}_i = \mathrm{tr}(\boldsymbol{S}_{D(i)})\quad \tilde{\lambda}_{1(i)}\ (i = 1, 2)$ で与える．次の検定統計量を考える．

$$F_1 = \frac{\tilde{\lambda}_{1(1)}}{\tilde{\lambda}_{1(2)}} \tilde{h}_* \tilde{\gamma}_*$$

ここで，$\tilde{h} = \max\{|\tilde{\boldsymbol{h}}_{1(1)}^T \tilde{\boldsymbol{h}}_{1(2)}|, |\tilde{\boldsymbol{h}}_{1(1)}^T \tilde{\boldsymbol{h}}_{1(2)}|^{-1}\}$, $\tilde{\gamma} = \max\{\tilde{\kappa}_1/\tilde{\kappa}_2, \tilde{\kappa}_2/\tilde{\kappa}_1\}$ とし，

$$(\tilde{h}_*, \tilde{\gamma}_*) = \begin{cases} (\tilde{h}, \tilde{\gamma}) & (\tilde{\lambda}_{1(1)} \quad \tilde{\lambda}_{1(2)} \text{のとき}), \\ (1/\tilde{h}, 1/\tilde{\gamma}) & (\tilde{\lambda}_{1(1)} < \tilde{\lambda}_{1(2)} \text{のとき}) \end{cases}$$

と定義する．そのとき，次の定理を得る．

定理 4．　適切な正則条件のもと，$p \to \infty$ のとき次が成り立つ．

$$F_1 \Rightarrow F_{n_1-1,n_2-1} \text{ under } H_0.$$

本報告では，提案した検定について，高い検出力を有することを理論的かつ数値的に確認した．また，マイクロアレイデータを用いた実データ解析も行った．提案した検定により，先行研究では見出すことのできなかった 2 つの共分散行列の差異を検出することができた．

参考文献

[1] Aoshima, M. and Yata, K. (2011). Two-stage procedures for high-dimensional data. *Sequential Anal. (Editor's special invited paper)* 30, 356-399.

[2] Ishii, A., Yata, K. and Aoshima, M. (2015). Asymptotic properties of the first principal component and equality tests of covariance matrices in high-dimension, low-sample-size context. Revised in *J. Stat. Plan. Inference*.

[3] Yata, K. and Aoshima, M. (2012). Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. *J. Multivariate Anal.* 105, 193-215.

漸近展開の不連続項を利用した
離散バートレット型変換統計量の性質について

北海道教育大・釧路　関谷祐里

鹿児島大・理工　種市信裕

1 多項分布の適合度検定における対数尤度比統計量の分布の近似

$(X_1,\ldots,X_k)'$ を多項分布 $M_k(n,\boldsymbol{\pi})$ に従う確率変数ベクトルとする。$\boldsymbol{p}=(p_1,\ldots,p_k)'$ を $0<p_j<1$ $(j=1,\ldots,k)$ と $\sum_{j=1}^{k}p_j=1$ を満たすある与えられたベクトルとするとき，単純帰無仮説

$$H_0:\boldsymbol{\pi}=\boldsymbol{p}$$

を検定するための対数尤度比統計量は

$$T=2\sum_{j=1}^{k}X_j\log\left(\frac{X_j}{np_j}\right)$$

で与えられ，H_0 のもとでの T の極限分布は，自由度 $k-1$ のカイ二乗分布である。多項分布 $M_k(n,\boldsymbol{p})$ に従う確率変数を $(X_1,\ldots,X_k)'$ とし，

$$Y_j=\frac{X_j-np_j}{\sqrt{n}}\qquad(j=1,\ldots,k),$$

$$\boldsymbol{Y}=(Y_1,\ldots,Y_r)',\qquad r=k-1$$

とおく。Siotani and Fujikoshi (1984) は，Yarnold (1972) の定理を適用することにより，H_0 のもとでの T の分布の下側確率に対する漸近展開式を以下のように与えた。

$$\Pr\{T<c|H_0\}=J_1+J_2+J_3+O(n^{-3/2})$$

ただし，J_1 は多変量エッジワース展開であり，

$$J_1=\Pr\{\chi^2_r<c\}+\frac{1}{n}\sum_{j=0}^{1}d_j\Pr\{\chi^2_{r+2j}<c\}+O(n^{-3/2})$$

と評価される。ここで，

$$d_0=\frac{1}{12}\left(1-\sum_{j=1}^{k}\frac{1}{p_j}\right),\quad d_1=-d_0$$

である。また，J_2 と J_3 は不連続項であり，J_2 項に対する漸近近似が，

$$\hat{J}_2 = Qn^{-r/2}\left\{N(c) - n^{r/2}V(c)\right\}$$

で与えられた。ここで，

$$Q = \left\{e^c(2\pi)^r \prod_{j=1}^{k} p_j\right\}^{-1/2},$$

$N(c)$ は集合 $B(c) = \{\boldsymbol{y} = (y_1, \ldots, y_r)' : T(\boldsymbol{y}) < c\}$ に含まれる格子点の数,$V(c)$ は $B(c)$ の体積で，$V(c) = \hat{V}(c) + O(n^{-3/2})$ と評価される。ただし，

$$\hat{V}(c) = \frac{\left\{(\pi c)^r \prod_{j=1}^{k} p_j\right\}^{1/2}}{\Gamma\left(\frac{r}{2}+1\right)}\left\{1 - \frac{c}{24(k+1)n}(S + 3k^2 - 6k + 2)\right\}.$$

もう一つの不連続項である J_3 項は非常に複雑であることと $J_3 = O(n^{-1})$ であることから，Siotani and Fujikoshi (1984) は単純帰無仮説 H_0 のもとでの対数尤度比統計量 T の下側確率 $\Pr\{T < c|H_0\}$ に対する近似式として，$J_1 + \hat{J}_2$ を提案した。

2 不連続項 $\hat{J}_2$ を利用した離散バートレット型変換統計量

対数尤度比統計量 T の連続項 J_1 のみを用いた（連続）バートレット変換統計量は，

$$T^B = \left(1 + \frac{2d_0}{nr}\right)T$$

で与えられる。また，関谷･種市 (2015) は，不連続項 J_2 の近似式 $\hat{J}_2$ の主要項を $n^{-1}\sum_{j=0}^{1} b_j^*$ $\Pr\{\chi^2_{r+2j} < c\}$ として形式的に表現しバートレット型変換を施すことによって，対数尤度比統計量 T に基づく変換統計量として，連続項 J_1 と不連続項 $\hat{J}_2$ を利用した離散バートレット型変換統計量 T^* 及び T^{**} を紹介した。本報告では，T から T^* や T^{**} への変換によって極限カイ二乗分布による近似がどの程度改善されたかを数値計算により考察し，どのような場合に不連続項 $\hat{J}_2$ の効果が認められるかを考察した。

参考文献

[1] 関谷祐里・種市信裕 (2015). 日本統計学会誌, **45**(1), 1-17. （印刷中）.

[2] M. Siotani and Y. Fujikoshi, *Hiroshima Math. J.*, **14**, (1984), 115–124.

[3] J. K. Yarnold, *Ann. Math. Statist.*, **43**, (1972), 1566–1580.

二項反応における一般化線型モデルのリンク関数の拡張

鹿児島大学・理工　種市信裕
北海道教育大学・釧路　関谷祐里
数学利用研究所　外山淳

1　はじめに

一般化線型モデル（Nelder and Wedderburn [3]）は，random component（ランダム成分），linear predictor（線型予測子），link function（リンク関数）の 3 つの成分で構成される．本報告においては，確率変数 X が二項分布 $B(n,\pi)$ に従う時, ランダム成分 $Y = X/n$ である場合の新しいリンク関数の構築およびそれによるモデルの提案をおこなった.

2　二項反応の一般化線型モデルにおける対称型および非対称型リンク関数

二項反応の一般化線型モデルにおけるリンク関数として，正準リンク関数であるロジットリンク関数 $g(t) = \log\{t/(1-t)\},\ (0 < t < 1)$ を用いると，ロジスティック回帰モデル

$$g(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \boldsymbol{x}'\boldsymbol{\beta}$$

が得られる．ただし，$\boldsymbol{x} = (x_1, \ldots, x_k)'$ は共変量ベクトル，$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ はパラメータベクトルである．また，リンク関数として，標準正規分布の累積分布関数 $\Phi(\cdot)$ の逆関数であるプロビットリンク $g(t) = \Phi^{-1}(t)$ を用いると，プロビットモデル

$$g(\pi) = \Phi^{-1}(\pi) = \boldsymbol{x}'\boldsymbol{\beta}$$

が得られる．今, リンク関数を L とすると，ロジットリンクやプロビットリンクは，

$$L(t) = -L(1-t),$$

という関係が成り立つという意味で，$t = 0.5$ の回りで対称なリンク関数である．このリンク関数の対称性にそぐわない観測度数および共変量の場合には，モデルへの適合度は悪くなる．このような観測度数および共変量の場合に有効なのが，補対数対数（complementary log-log）リンクを用いた次の補対数対数モデルである.

$$g(\pi) = \log\{-\log(1-\pi)\} = \boldsymbol{x}'\boldsymbol{\beta}$$

Aranda-Ordaz [1] は，対称性を持つロジットリンクから非対称な補対数対数リンクまでを包括的に扱うために，パラメータ c に依存するリンク関数

$$g_c^A(t) = \log\left\{\frac{(1-t)^{-c} - 1}{c}\right\} \tag{1}$$

の族を提案した．このリンク関数は，$c = 1$ のときロジットリンクとなり，$c \to 0$ の時極限として補対数対数リンクに一致する．

３　新たなリンク関数族の構築と適用の基準およびデータへの適用

本報告では，ロジットリンク関数や補対数対数リンク関数はもちろん，Aranda-Ordaz [1] によるリンク関数の族をすべて含むさらに大きなリンク関数の族の提案をおこなう．この新たなリンク関数族を考えることによって，従来のリンク関数ではうまく適合しなかったデータに対しても，適切なリンク関数に基づくモデルを見つけられる可能性が生まれる．本報告では, 以下のように提案をおこなった．

1. **導出法:** リンク関数は値域が [0,1] の狭義単調増加関数 F の逆関数で与えることができる。F としては累積分布関数 (CDF) を考えることが適当であり, 通常おこなわれている。累積分布関数の逆関数で与えられたリンクは逆 CDF リンクと呼ばれ, ロジットリンク, プロビットリンク, 補対数対数リンクはすべて逆 CDF リンクである。ロジットリンク関数や補対数対数リンク関数はもちろん, Aranda-Ordaz によるリンク関数をすべて含んだ, 累積分布関数の逆関数を用いたリンク関数（逆 CDF リンク）による新しいタイプのリンク関数族の構築をおこなった。

2. **歪度, 尖度によるリンク関数の効用の可能性確認:** リンク関数の違いや性質を調べるために, パラメータの変化によってリンク関数の逆関数 $F(x)$ が示す分布の対称性を表現する歪度と, 裾の厚さを表現する尖度を指標として用い様々なデータに対応できるかの確認を行った。その結果, 提案されたリンク関数の逆関数の累積分布関数には, Aranda-Ordaz によるリンク関数を与える累積分布関数では達成できない歪度と尖度の組の範囲があることを示した．

3. **モデルの適合度検定:** Aranda-Ordaz によるモデルで十分なのか提案されたモデルが必要なのかを評価する検定方法の提示をおこなった．

4. **データへの検定の適用:** Aranda-Ordaz によるモデルに適合しにくい特徴を持つ人工データおよびリアルデータへの検定の適用をおこない, 実際のデータに対して提案されたモデル族の必要性を示した．

参考文献

[1] Aranda-Ordaz, F. J.: On two families of transformations to additivity for binary response data. *Biometrika*, **68**, (1981), 357–363.

[2] Nelder, J. A. and Wedderburn, R. W. M.: Generalized linear models. *J. R. Statist. Soc. A*, **135**, (1972), 370–384.