科学研究費・基盤研究(A)「非対称・非線形統計理論と経済・生体科学 への応用」（代表：谷口正信（早稲田大学基幹理工学部））によるシンポジウム

「一般化線形モデルの最新の展開とその周辺」

■ 開催日時：2013 年 11 月 8 日（金）～11 月 10 日（日）
■ 会場：けやき会館（千葉大学西千葉キャンバス）　会議室４

目　　次

変量効果モデルを用いた
複数の読影者による画像診断法の精度の推定

佐伯　浩之 [1,2]、汪 金芳 [1]

[1] 千葉大学大学院 理学研究科
[2] 富士フイルムＲＩファーマ株式会社

1　目的

ある疾患を検査する診断法の精度を示す指標として、感度（Sensitivity）と特異度（Specificity）がある。一般に、感度とは疾患を持つ被験者に対して診断法が陽性 (+) を示す確率、特異度とは疾患を持たない被験者に対して診断法が陰性 (−) を示す確率と定義される。画像診断法の精度を推定するための臨床試験では、読影者の評価の再現性を確認するために、複数の読影者が同一の画像を独立に読影する。従って、複数の読影者により複数の判定が発生することから、従来は合議や多数決によって判定を単一化した上で、感度や特異度を推定していた。しかしながら、合議では非独立な評価のためにバイアスが発生する恐れがあること、多数決評価では読影者間のバラツキを考慮できないことから、これら方法の利用は主要な評価に対して推奨されない。そこで本研究では、複数の読影者から得られた診断法の精度の推定値を統合する方法を導出することを目的とした。

2　統計モデル

ある疾患 D を検査する診断法 M の性能を示す指標である感度 p と特異度 q を以下のように定義する。

$$p = P(+|D) \tag{1}$$

$$q = P(-|\bar{D}) \tag{2}$$

ここで、$\bar{D}$ は疾患ではない状況を意味する。読影者母集団 $\mathcal{R}$ から、無作為に K 名の読影者を抽出し、$R_1, \cdots, R_K$ とする。読影者 R_k が画像を読影し、Y_k を出力する。ここで、読影者 R_k が陽性と判定した場合には、$Y_k = +$ となり、陰性と判定した場合には、$Y_k = -$ となる。したがって、Y_k はベルヌーイ分布に従い、陽性及び陰性判定の確率変数として以下のように表せる。

$$p_k = P(Y_k = +|D) = P(+|D, R_k)\,, \quad k = 1, \cdots, K \tag{3}$$

$$q_k = P(Y_k = -|\bar{D}) = P(-|\bar{D}, R_k)\,, \quad k = 1, \cdots, K \tag{4}$$

以降では、最も単純な状況として読影者 2 名での判定結果に基く感度を対象として検討を進める。読影者 1 及び 2 が同一の画像を読影した結果を y_1 及び y_2 とし、以下のモデルを定義する。

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} \tag{5}$$

ここで、θ_1 と θ_2 は読影者 1 及び 2 の読影結果の平均である。また、ϵ_1 及び ϵ_2 は、分散が $\hat{s}_1^2$ と $\hat{s}_2^2$（既知）、同一画像の読影による相関を考慮した相関係数が ρ_{12} の多変量正規分布に従うと仮定する。

$$\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} \ , \ \begin{pmatrix} \hat{s}_1^2 & \rho_{12}\hat{s}_1\hat{s}_2 \\ \rho_{12}\hat{s}_1\hat{s}_2 & \hat{s}_2^2 \end{pmatrix} \right) \tag{6}$$

更に、θ_1 と θ_2 に対して以下のモデルを定義する。

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} \mu \\ \mu \end{pmatrix} + \begin{pmatrix} e \\ e \end{pmatrix} \tag{7}$$

ここで、μ は全ての読影者での読影結果の平均である。また、e は分散が τ^2（未知）の多変量正規分布に従うと仮定する。

$$\begin{pmatrix} e \\ e \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} \ , \ \begin{pmatrix} \tau^2 & 0 \\ 0 & \tau^2 \end{pmatrix} \right) \tag{8}$$

以上で定義したモデルの式 (5) 及び (7) に基づき、以下の random effects model を定義する。

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \mu \\ \mu \end{pmatrix} + \begin{pmatrix} \epsilon_1^* \\ \epsilon_2^* \end{pmatrix} \tag{9}$$

$$\begin{pmatrix} \epsilon_1^* \\ \epsilon_2^* \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} \ , \ \begin{pmatrix} \hat{s}_1^2 + \tau^2 & \rho_{12}\hat{s}_1\hat{s}_2 \\ \rho_{12}\hat{s}_1\hat{s}_2 & \hat{s}_2^2 + \tau^2 \end{pmatrix} \right) \tag{10}$$

本研究のアウトカムである読影者 1 及び 2 の感度 $\hat{p}_1$ 及び $\hat{p}_2$ は割合であることから、ロジット変換を行った上で式 (9) を適応することとなる。その際のモデル式と誤差は以下のように示される [1]。

$$\begin{pmatrix} \mathrm{logit}(\hat{p}_1) \\ \mathrm{logit}(\hat{p}_2) \end{pmatrix} = \begin{pmatrix} \mathrm{logit}(p) \\ \mathrm{logit}(p) \end{pmatrix} + \begin{pmatrix} \epsilon_1^* \\ \epsilon_2^* \end{pmatrix} \tag{11}$$

$$\begin{pmatrix} \epsilon_1^* \\ \epsilon_2^* \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} \ , \ \begin{pmatrix} \frac{1}{n\hat{p}_1(1-\hat{p}_1)} + \tau^2 & -\frac{1}{n(1-\hat{p}_1)(1-\hat{p}_2)} \\ -\frac{1}{n(1-\hat{p}_1)(1-\hat{p}_2)} & \frac{1}{n\hat{p}_2(1-\hat{p}_2)} + \tau^2 \end{pmatrix} \right) \tag{12}$$

3 精度の推定

μ の推定量として単純平均値 $\hat{\mu}_{(MEAN)}$、最尤推定量 $\hat{\mu}_{(AMLE)}$ 及び制限付き最尤推定量 $\hat{\mu}_{(REML)}$ を、以下の式により求める。

$$\hat{\mu}_{(MEAN)} = \left(X^t X\right)^{-1} X^t \mathbf{Y} \tag{13}$$

$$\hat{\mu}_{(AMLE)} = \left(X^t \hat{R}^{-1} X\right)^{-1} X^t \hat{R}^{-1} \mathbf{Y} \tag{14}$$

$$\hat{\mu}_{(REML)} = \left(X^t \hat{V}^{-1} X\right)^{-1} X^t \hat{V}^{-1} \mathbf{Y} \tag{15}$$

ここで、

$$X = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$\hat{R} = \begin{pmatrix} \hat{s}_1^2 & \rho_{12}\hat{s}_1\hat{s}_2 \\ \rho_{12}\hat{s}_1\hat{s}_2 & \hat{s}_2^2 \end{pmatrix}$$

$$\hat{V} = \begin{pmatrix} \hat{s}_1^2 + \hat{\tau}^2 & \rho_{12}\hat{s}_1\hat{s}_2 \\ \rho_{12}\hat{s}_1\hat{s}_2 & \hat{s}_2^2 + \hat{\tau}^2 \end{pmatrix}$$

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

である。本研究のアウトカムである感度 $\hat{p}_{(MEAN)}$、$\hat{p}_{(AMLE)}$ 及び $\hat{p}_{(REML)}$ は、ロジット変換に基いた $\hat{\mu}_{(MEAN)}$、$\hat{\mu}_{(AMLE)}$ 及び $\hat{\mu}_{(REML)}$ を逆変換することで求める。なお、以下の式の $\hat{p}_{(.)}$ 及び $\hat{\mu}_{(.)}$ は、添え字を省略したものである。

$$\hat{p}_{(.)} = \frac{exp\left(\hat{\mu}_{(.)}\right)}{1 + exp\left(\hat{\mu}_{(.)}\right)} \tag{16}$$

なお、τ^2 の推定量である $\hat{\tau}^2$ は、DerSimonian and Laird の方法によりモーメント推定量として求めることができる [2]。

$$Q_1 = X^t \hat{R}^{-1} \left(\mathbf{Y} - \begin{pmatrix} \mu_{(AMLE)} \\ \mu_{(AMLE)} \end{pmatrix} \right) \left(\mathbf{Y} - \begin{pmatrix} \mu_{(AMLE)} \\ \mu_{(AMLE)} \end{pmatrix} \right)^t X \tag{17}$$

$$\hat{\tau}^2 = max \left[0, \{Q_1 - (K-1)\} \left\{ X^t \hat{R}^{-1} X - \left(X^t \hat{R}^{-1} \right) \left(X^t \hat{R}^{-1} \right)^t \left(X^t \hat{R}^{-1} X \right)^{-1} \right\}^{-1} \right] \tag{18}$$

4 数値実験

本研究で誘導した $\hat{p}_{(MEAN)}$、$\hat{p}_{(AMLE)}$ 及び $\hat{p}_{(REML)}$ のバイアスと MSE を検討するため、2 名の読影者の条件のもとでモンテカルロシミュレーションを実施した。シミュレーションデータは 2 段階の過程を経て作成した。まず最初に、任意の p (シミュレーションではロジット変換する)、$\hat{s}_1^2$、$\hat{s}_2^2$、ρ_{12} 及び τ^2 について任意の値を設定したもとで多変量正規乱数を発生させ、p_1 及び p_2 の擬似データを作成した。次に、この擬似データから p_1 及び p_2 のもとで任意の n 数のベルヌーイ乱数による擬似データを発生させた。シミュレーションの結果、$\hat{p}_{(AMLE)}$ は $\hat{p}_{(MEAN)}$ 及び $\hat{p}_{(REMS)}$ に対してバイアスと MSE が大きかった。$\hat{p}_{(MEAN)}$ 及び $\hat{p}_{(REMS)}$ を比べると、$\hat{p}_{(MEAN)}$ が若干バイアスが小さい傾向が認めれらたが、その差は小さかった。

参考文献

[1] Trikalinos T, Olkin I. A method for the meta-analysis of mutually exclusive binary outcomes. *Statistics in Medicine* 2008; **27**:4279–4300.

[2] DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**:177–188.

Determination of smoothing parameter for penalized spline regression

吉田拓真 [1]
[1] 鹿児島大学　大学院理工学研究科

1　はじめに

スプライン平滑化において, 推定量の挙動を制御する平滑化パラメータの設定は重要である. そのための代表的な手法として, cross-validation などのモデル選択規準を利用した探索決定法がある. しかし, 探索決定法はどうしても計算時間がかかり, 複数の平滑化パラメータの候補の中から最適なものをひとつ選ぶという特徴から, すべての候補点が悪いということさえ起こり得る. 決定すべき平滑化パラメータが 1 つであれば, 近年の計算機の性能向上によってその欠点はカバーできる. しかし, 決定すべき平滑化パラメータが複数存在する場合は, 膨大な計算コストがかかり, 利便性があるとは言えない. もし, 推定量が持つ数学的性質から平滑化パラメータをダイレクトに決定することが可能であれば, 計算時間の短縮だけでなく, 数学的な理論保証を持つ推定量が構築できる.

本講演では, スプライン推定量の平均積分二乗誤差最小化に基づいて平滑化パラメータを決定する手法に関するアルゴリズム, 理論結果, 適用結果を報告した. 提案した手法は説明変数, 目的変数が共に 1 次元の場合のみならず, 説明変数が多次元の加法モデルに対しても適用可能である. 加法モデルにおけるスプライン平滑化を考える場合は, 平滑化パラメータは説明変数の個数だけ決定する必要がある. 提案手法を用いると従来の探索的な決定よりも格段に少ない計算時間で推定量の構築が可能となった.

2　スプライン推定量

得られたデータ $\{(y_i, x_i) : i = 1, \cdots, n\}$ に対して, 回帰モデル

$$Y_i = f(x_i) + \varepsilon_i, \quad i = 1, \cdots, n, \tag{1}$$

を考える. ここで, Y_i は目的変数, x_i は説明変数, f は未知の回帰関数, ε_i は誤差であり, $E[\varepsilon_i] = 0$, $V[\varepsilon_i] = \sigma^2 < \infty$ と仮定する. 目的は, f をスプライン法によって推定することである. 節点 $\kappa_0 < \kappa_1 < \cdots < \kappa_K < \kappa_{K+1}$ に対して, p 次の B-スプライン関数を, $k = -p, \cdots, K-1$ として

$$\begin{aligned} B_k^{[0]}(x) &= \begin{cases} 1, & \kappa_k < x \le \kappa_{k+1}, \\ 0, & \text{otherwise}, \end{cases} \\ B_k^{[p]}(x) &= \frac{x - \kappa_k}{\kappa_{k+p} - \kappa_k} B_k^{[p-1]}(x) + \frac{\kappa_{k+p+1} - x}{\kappa_{k+p+1} - \kappa_{k+1}} B_{k+1}^{[p-1]}(x) \end{aligned}$$

と定義する. 以降は, $B_k(x) = B_k^{[p]}(x)$ と書く. このとき, 回帰モデル (1) の代わりに, B-スプライン回帰モデル

$$Y_i = s(x_i) + \varepsilon_i, \quad i = 1, \cdots, n,$$

を考える. ただし,

$$s(x) = \sum_{k=-p}^{K-1} B_k(x) b_k$$

である．目的は, パラメータ $\boldsymbol{b} = (b_{-p} \ \cdots \ b_{K-1})^T$ を推定することである．罰則付きスプライン $\hat{\boldsymbol{b}} = (\hat{b}_{-p} \ \cdots \ \hat{b}_{K-1})^T$ は

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{k=-p}^{K-1} B_k(x) b_k \right\}^2 + \lambda \sum_{k=m-p}^{K-1} \{\Delta^m(b_k)\}^2, \tag{2}$$

の最小化に基づき決定される．ここで, λ は平滑化パラメータ, Δ は差分作用素であり, $\Delta b_k = b_k - b_{k-1}$ で定義される．このとき, f の推定量は $\hat{f}(x) = \sum_{k=-p+1}^{K} B_k(x)\hat{b}_k$ で構成される．

3　平滑化パラメータの決定法

$\hat{f}(x)$ の平均積分二乗誤差は

$$\mathrm{MISE} = \int E[\{\hat{f}(x) - f(x)\}^2]dx$$

で定義される．この MISE を最小にする λ をひとつ決定する．漸近表現用いると, MISE は λ に関して 2 次関数の形で書ける．つまり, 定数 a, b, c を用いて,

$$\mathrm{MISE} \approx a\lambda^2 + b\lambda + c \tag{3}$$

の形で書ける．よって, MISE を最小にする λ は簡単に求めることができる．この手法のポイントは, (3) を満たす a, b, c を求めることであり, そのためにテイラー展開, B-スプラインの階層的な微分計算, 逆行列の漸近展開など数学的な厳密な計算が求められた.

4　加法モデルへの拡張

以上の議論を加法モデルへ拡張する．加法モデルとは, 1 次元の目的変数 Y_i と d 次元説明変数 $(x_{i1}, \cdots, x_{id})$ について

$$Y_i = \mu + f_1(x_{i1}) + \cdots + f_d(x_{id}) + \varepsilon_i. \tag{4}$$

と仮定することである．μ は未知のパラメータであり, f_j は未知の回帰関数である．加法モデル (4) に対してスプライン法を適用する場合は, 各 f_j に対してスプラインモデルで近似し, (2) に基づいて f_j を推定するため, 決定が必要となる平滑化パラメータは $\lambda_1, \cdots, \lambda_d$ の d 個となる．このモデルに関しても各 f_j の罰則付きスプライン $\hat{f}_j$ について

$$\mathrm{MISE}_j = \int E[\{\hat{f}_j(x) - f_j(x)\}^2]dx$$

を最小にする λ_j を決定する．すると, (3) と同様に, MISE_j は漸近的に λ_j の 2 次関数で書けるので, これを求めた.

5　数値実験

シンポジウムにおいて, 提案手法の有用性を数値的に示し, 報告した.

Functional clustering of mouse ultrasonic vocalization data

Xiaoling Dou **The Institute of Statistical Mathematics**
Shingo Shirahata **Osaka University**
Hiroki Sugimoto **Jichi Medical University**
Tsuyoshi Koide **National Institute of Genetics**

1 Introduction

Mouse ultrasonic vocalizations (USVs) are studied in various fields of science ([3], [4]). However, background noise and varied USV patterns in observed signals make complete automatic analysis difficult. We propose a series methods to cluster nonharmonic mouse USV data automatically. The procedure includes noise reduction, detecting USV calls, transforming USV calls as functions and functional clustering ([1], [2]). The proposed methods are shown useful with two data sets taken from laboratory mice.

2 Mouse ultrasonic vocalization data

Our USV data are taken from male mice who emit towards a female mouse. Two strains, BALB/cAnN and C57BL/6, are used in the experiments. For the dataset taken from a mouse of BALB/cAnN, the raw USV voice data are transformed by FFT into a sequence of USV frames. Each frame of 0.85 msec duration is observed. Each USV frame overlaps 50% with the next frame, and each frame is represented by 256 samples at a 300kHz sampling rate. That is, the measurable highest frequency (the Nyquist frequency) is 150kHz which is half the sample frequency, and the time between adjacent frames is 0.425 msec. For each frame, 126 squared amplitudes (or power) are computed for the sampled frequencies. Thus, from the BALB/cAnN mouse, the data collected from about 100 seconds consists the squared amplitudes as a 233393×126 matrix.

The second dataset is taken from a male C57BL/6 mouse. It contains about 188 seconds USV data. After the FFT described above, it became a squared amplitude matrix with size 443086×126. Compared with those of BALB/cAnN, most USV calls of the C57BL/6 mouse are discontinuous.

3 Noise reduction

There are three steps in the noise reduction process: moving average, determine the minimum USV intensity and separate the USV signals from the matrix.

We denote the matrix by $X = \{x_{ij}\}$, $i = 1, ..., T$ stands for time, $j = 1, ..., F$ indicates frequency. First we use a small matrix to do moving average on the original matrix data. The size of the small matrix can be chosen empirically. Moving average is useful to distinguish noise and voice signal. It can also reduce the perturbation in the signals.

The second step is to determine a minimum intensity (e.g., 200) and set all elements less than the intensity to 0. This minimum intensity should be chosen by considering the noise level. This step etches the shapes of the USV signals. We then find the location in the frequency range where the signal takes its maximum intensity for each frame. If more than one locations are found for one frame, the median should be used. Generally, frequency of USV is higher than 20KHz, the location should be chosen at larger than 20KHz. In this paper, we only consider the shape of the USVs in time, but not the intensity which measures USV volume. Now, we have a sequence contains the frequencies of the USVs in time. That is, the large matrix becomes a two-dimensional data, one dimension is for time, the other one is for frequency.

We can then separate the USV calls from the sequence, after determine the length of interval between adjacent USVs and the range of the USV length.

4 Function approximation and functional clustering

After reducing the noise, we are now ready to define the USV calls as functions. Many methods are available for estimating the USV functions. Considering that some of the USV calls are not continuous, they may contain several breakpoints, we prefer the B-spline basis function method for its easiness in constructing functions with breakpoints.

For n pairs of (t_i, f_i), $i = 1, \cdots, n$, contained in one USV call. Assume that

$$f_i = \theta_0 + \sum_{j=1}^{m} \theta_j \beta_j(t_i) + \varepsilon_i \tag{1}$$

or in a matrix form

$$\boldsymbol{f} = \boldsymbol{B}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{2}$$

where $\boldsymbol{t} = (t_1, \cdots, t_n)'$, $\boldsymbol{f} = (f_1, \cdots, f_n)'$ stand for time and frequency, respectively. $\boldsymbol{B} = (\boldsymbol{1}, \beta_1(t), \cdots, \beta_m(t))$ is the B-spline basis functions with order d. Using the least square method, for each USV call, we obtain a coefficient vector

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{f} \tag{3}$$

and the regression function

$$\widehat{\boldsymbol{f}} = \boldsymbol{B}\widehat{\boldsymbol{\theta}}. \tag{4}$$

For discontinuous USV calls, we specify a constant κ as threshold. If $|f_i - f_{i-1}| > \kappa$, then the curve is discrete at time i, and i is called a breakpoint. To make a jump at a breakpoint, we overlap $d + 1$ knots at the breakpoint. Hence, for each curve, we obtain a coefficient vector with $m = d +$ number of interior knots.

Because curves with same number of breakpoints have the same length of coefficient vectors, clustering can be performed by first grouping the curves with same number of breakpoints into the first level clusters. Then the shape of the curves in each first level cluster are classified by clustering their coefficient vectors ([1]). In the step of functional clustering, we use Ward's method and K-means method.

5 Data analysis

After noise reduction, we obtained 390 USV calls form the data of the BALB/cAnN mouse. Among them, 378 curves are continuous, and are clustered into 5 groups. The rest 12 curves with one breakpoint are gathered into one cluster. These methods also work well for the data taken from the C57BL/6 mouse.

References

[1] Abraham, C., Cornillon, P. A., Matzner-Lober, E., and Molinari, N. (2003). Unsupervised curve clustering using B-splines. *Scandinavian Journal of Statistics*, **30**, 581–595.

[2] Ramsay, J. O. , Silverman, B. W. (2006). *Functional Data Analysis*, 2nd ed., Springer.

[3] Scattoni, M. L., Gandhy, S. U., Ricceri, L., and Crawley, J. N. (2008). Unusual repertoire of vocalizations in the BTBR T+tf/J mouse model of autism, *Plos ONE*, **3**(8), e3067.

[4] Sugimoto, H., Okabe, S., Kato, M., Koshida, N., Shiroishi, T., Mogi, K., Kikusui, T., and Koide, T. (2011). A Role for Strain Differences in Waveforms of Ultrasonic Vocalization during Male-Female Interaction, *PLoS ONE*, **6**(7), e22093.

チューブの体積を最小にするフーリエ・多項式回帰最適実験計画

栗木哲 (統計数理研)　Henry Wynn (LSE, UK)

1　最適実験計画とは

実験データ $(x_i, y_i)_{i=1,\dots,N}$ に回帰分析モデル

$$y_i = b^\top f(x_i) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_0^2) \ \text{ i.i.d.}$$

を想定する．ここで $b = (b_1, \dots, b_n)^\top$ は未知パラメータベクトル，$f(x) = (f_1(x), \dots, f_n(x))^\top$ は基底ベクトルである．

説明変数 x_i は，その定義域 $\mathcal{X} \subset \mathbb{R}$ の中で自由に設定することができるとする．未知の b を何らかの意味で精度良く推定するための $\{x_1, \dots, x_n\} \subset \mathcal{X}$ の配置を求める問題を最適実験計画という．b の最小 2 乗推定量 $\widehat{b}$ の共分散行列は $\sigma_0^2\Sigma$,

$$\Sigma = M^{-1}, \quad M = \textstyle\sum_{i=1}^N f(x_i) f(x_i)^\top$$

であるので，$\det(\Sigma)$ を最小にする D 最適計画，$\mathrm{tr}(\Sigma)$ を最小にする A 最適計画，$\widehat{b}^\top f(x_0)$ の分散 $f(x_0)^\top \Sigma f(x_0)$ を最小にする計画などが考えられる．

2　チューブ法による同時信頼区間

回帰曲線 $\{(x, b^\top f(x)) \mid x \in \mathcal{X}\}$ の信頼区間を同時信頼区間という．それを構成するための簡単な方法は Cauchy-Schwarz 不等式によるものであるが，特別な場合を除いて，得られる不等式はタイトではない．

$\mathbb{R}^n$ の計量を $\|u\|_\Sigma^2 = u^\top \Sigma u$ で考える．半径 1 の球面 $\mathbb{S}^{n-1}$ の上の曲線 $\Gamma = \{f(x)/\|f(x)\|_\Sigma \mid x \in \mathcal{X}\}$ を考える．D. Naiman (1986) は，曲線 Γ の $\mathbb{S}^{n-1}$ における管状近傍 (チューブ) の体積を考察することにより，確率不等式

$$1 - \Pr\left(\frac{|\widehat{b}^\top f(x) - b^\top f(x)|}{\|f(x)\|_\Sigma} < c, \forall x \in \mathcal{X} \right) \le \mathrm{length}(\Gamma) \Pr\big(\chi_2^2 > c^2\big) + \chi(\Gamma) \Pr\big(\chi_1^2 > c^2\big) \tag{1}$$

を示した．ここで $\mathrm{length}(\Gamma)$ は Γ の長さ，$\chi(\Gamma)$ は Γ の連結成分数である．右辺を α としておいて c について解くことにより，信頼係数 $1-\alpha$ の同時信頼区間が構成できる．(1) は (安全側の) 不等式であるが，裾領域 $c \to \infty$ $(\alpha \to 0)$ では左辺と右辺は漸近的に同等となる．右辺は第 1 項が主要項となる．

3　問題設定

(1) より，$\|f(x)\|_\Sigma$ を小さくする，また曲線の長さ $\mathrm{length}(\Gamma)$ を短くすると同時信頼区間の幅を狭くすることが分かる．前者を $\min \max_{x \in \mathcal{X}} \|f(x)\|_\Sigma$ の意味で達成するのが D 最適計画である．ここでは $\mathrm{length}(\Gamma)$ を最小にする最適問題を考える．

特に具体例として，基底を

$$f(x) = \big(1, \sqrt{2}\sin(2\pi x), \sqrt{2}\cos(2\pi x), \sqrt{2}\sin(4\pi x), \dots, \sqrt{2}\cos(2\pi m x)\big)^\top$$

$(n = 2m + 1)$ とするフーリエ回帰を考える．$\mathcal{X} = (-1/2, 1/2]$ とする．

$\mathcal{X}$ 上の非負測度の全体を $\mathcal{P}$ で表す．$P \in \mathcal{P}$ のそれぞれが実験配置に対応する．正則な情報行列の全体 (モーメント錐) を $\mathcal{M} = \left\{\int_{\mathcal{X}} f(x)f(x)^\top dP(x) \succ 0 \mid P \in \mathcal{P}\right\}$ と書く．$g(x) = \frac{d}{dx}f(x)$ とおく．我々の問題は次の形に帰着される．

`Minimize`

$$\mathrm{length}(\Gamma) = \int_{\mathcal{X}} \frac{\sqrt{(f(x)M^{-1}f(x))(g(x)M^{-1}g(x)) - (f(x)M^{-1}g(x))^2}}{f(x)^\top M^{-1} f(x)}\, dx \quad (2)$$

`subject to` $M \in \mathcal{M}$.

4　問題解決への道のり

以下のことがらを見いだし，$n = 3$ の場合を解決した．

(i) フーリエ回帰における一様計画の場合 (2) は積分可能で $\mathrm{length}(\Gamma) = 2\pi\sqrt{2/3}$，局所最小解．(当初これが最適解と予想された．)

(ii) D 最適計画，A 最適計画とは異なり，目的関数 $\mathrm{length}(\Gamma)$ は $M = \Sigma^{-1}$ の凸関数ではなく，凸最適化の標準的な手法が使えない．

(iii) 基底を $f(x) = (1, x, \ldots, x^{n-1})^\top$ とする多項式回帰モデルでも同じ問題を考えることができる．変換 $x \mapsto \tan(\pi x)$ によって，多項式回帰での最適解とフーリエ回帰での最適解は 1 対 1 に対応する．

(iv) (2) の被積分関数の分子分母の多項式の終結式を計算し共通因子をもつ場合を調べたところ，その場合に平方根が消え (2) は積分可能となり，さらに積分値はフーリエ一様計画の場合に一致した．(最適解は一点ではなく代数多様体をなすと予想できた．)

(v) フーリエ回帰における角度シフトを一般化した変換

$$x \mapsto \frac{ax + b}{cx + d} \quad (ad - bc \neq 0)$$

(メビウス変換) は問題を不変にする．

(vi) メビウス変換によって，モーメント錐 $\mathcal{M}$ はオービット分解される．オービットの代表元として

$$M = \begin{pmatrix} 1 & 0 & v \\ 0 & v & 0 \\ v & 0 & 1 \end{pmatrix} \in \mathcal{M}, \quad v \in (0, 1/3]$$

をとることができる (クロスセクション)．$\mathrm{length}(\Gamma)$ の最小化問題は 1 次元パラメータ v に関する最適化問題に帰着する．しかしまだそれでも楕円積分が必要．

(vii) 不等式 $1/\sqrt{1+a} \geq 1 - a/2$, $\sqrt{1+a} \leq 1 + a/2$ $(|a| \leq 1)$ を用いて，$\mathrm{length}(\Gamma)$ の楕円積分によらない上下限を得ることができる．この上下限の最適化を通して最適解 $(v = 1/3)$ が得られた！

定理　$\mathrm{length}(\Gamma)$ の最小値は，多項式回帰のモーメント行列が

$$M = \begin{pmatrix} 1 & r & \frac{q^2}{3} + r^2 \\ r & \frac{q^2}{3} + r^2 & r(q^2 + r^2) \\ \frac{q^2}{3} + r^2 & r(q^2 + r^2) & (q^2 + r^2)^2 \end{pmatrix}, \quad q \neq 0$$

のとき達成される．$(q, r) = (1, 0)$ の場合は，フーリエ回帰の一様計画 (D 最適) に対応する．

海洋生物多様性の評価方法

–混合効果モデルとブースティングの応用–

小森　理
統計数理研究所

1　報告内容

海洋資源評価の問題は，資源の適切な管理運用にとって重要な問題である．(Worm *and others*, 2006) の論文では現状の資源管理体制のままでは 2048 年までに漁業資源の崩壊が示唆されており，海洋資源の評価，管理の見直しの動きが活発になってきている．このような状況の中，近年では個々の stock に注目した資源の保全ではなく，世界規模の資源管理が注目を浴びるようになってきた (Thorson *and others*, 2012; Costello *and others*, 2012)．世界規模のデータは食糧農業機関 (FAO) に蓄積されている漁獲量のデータのみであり，海洋資源評価の基準となる biomass の情報はない．そこで小規模ではあるが biomass の情報もある RAM データを使い，漁獲量と biomass とを関係づけるモデルの構築が必要となる．今回は時系列データである漁獲量のパターン抽出のために関数ロジスティックモデルを考え，周辺尤度最大化による手法とブースティングによる手法で判別解析を試みた．

2　今後

水産資源の豊かさの評価を難しくしている要因は，評価に利用できるデータが不足しているまたはそのデータの質が不確かな点があげられる．水産資源の評価で一番信頼性のある biomass の算出 (stock assessments) には多額の費用がかかり，全ての stock に対して算出することは現実的ではない．上記のようにごく少数の stock にのみこの biomass のデータが算出されている (RAM データ)．そこで実質的には FAO の漁獲量のデータが注目されているが注意すべき点がいくつか列挙される．まず漁獲量自体均質な量ではないことである．例としてカナダのタラ資源を考えてみる．この資源は以前から慎重に管理されてきたが 1990 年ごろに collapse してしまった．原因は trawler による漁獲量のみに注目してしまい，その他の boat 等を使った際の漁獲量の変動には注意を払わなかったからである．また漁獲量のデータ自体の信頼性の問題もある．発展途上国と先進国での漁獲量の報告量には明らかな違いが指摘されており，発展途上国での漁獲量が過小に報告されている事実がある．その他にも政府による漁獲の規制，stock の分類方法の再編，自然災害等のさまざまな要因が漁獲量のデータの信頼性を不確かなものにしている (Pauly *and others*, 2013)．このような事実を踏まえた上でデータ解析を行う必要がある．漁獲地域ごとのランダム効果を考えると

ともに，データのもつ不確かさをミスラベルモデル (Copas, 1988; Takenouchi and Eguchi, 2004) として捉える手法を考えたい．ミスラベルの割合の推定には OpenBugs 等を使ったベイズの手法が有効と考えている．

参考文献

Copas, J. (1988). Binary Regression Models for Contaminated Data. *Journal of the Royal Statistical Society: Series B.* **50**, 225–265.

Costello, C., Ovando, D., Hilborn, R., Gaines, S. D., Deschenes, O. and Lester, S. E. (2012). Status and Solutions for the World's Unassessed Fisheries. *Science* **338**, 517–520.

Pauly, D., Hilborn, R. and Branch, T. A. (2013). Fisheries: Does catch reflect abundance? *Nature* **494**, 303–306.

Takenouchi, T. and Eguchi, S. (2004). Robustifying AdaBoost by adding the naive error rate. *Neural Computation* **16**, 767–787.

Thorson, J. T., Branch, T. A. and Jensen, O. P. (2012). Using model-based inference to evaluate global fisheries status from landings, location, and life history data. *Can. J. Fish. Aquat. Sci* **69**, 645–655.

Worm, B., Barbier, E. B., Beaumont, N., Duffy, J. E., Folke, C., Halpern, B. S., Jackson, J. B. C., Lotze, H. K., Micheli, F., Palumbi, S. R., Sala, E., Selkoe, K. A., Stachowicz, J. J. and Watson, R. (2006). Impacts of biodiversity loss on ocean ecosystem services. *Science* **314**, 787–790.

潜在結果変数モデルに基づく直接効果・間接効果の推定

田栗正隆

taguri@yokohama-cu.ac.jp

横浜市立大学学術院医学群臨床統計学・疫学

〒236-0004 横浜市金沢区福浦 3-9

疫学研究の 1 つの目的は，興味のある曝露の疾病発生に対する因果効果を推定することである．曝露と疾病の間に因果関係が示唆された場合，どのようなメカニズムで効果があるのかについての知見を得ることにも興味が持たれる場合がある. この問題に対しての 1 つのアプローチは, 曝露の疾病に対する影響(総合効果)を，中間変数を介しない直接効果と，中間変数を介した間接効果に分解することである．本発表では，潜在結果変数モデルに基づく直接効果・間接効果の定義と識別のための仮定，識別式についてまとめを行った.

$Y(a)$と $M(a)$を，曝露変数 A が a という値に固定された場合に観察されたであろう潜在的な結果変数および中間変数の値とする．同様に，$Y(a,m)$を曝露変数 A が a，中間変数 M が m という値に固定された場合に観察されたであろう潜在的な結果変数の値とする．また，観察データと潜在データを関連付けるために，以下に述べる一致性の仮定(consistency assumption)と構成性の仮定(composition assumption)を置く．一致性の仮定は，$A = a$ かつ $M = m$ が観察されたサブグループでは，結果変数 Y が潜在結果変数 $Y(a,m)$と一致するという仮定である．同様に，$A = a$ が観察されたサブグループでは，中間変数 M が潜在変数 $M(a)$と一致することを仮定する．また，構成性の仮定は，$Y(a) = Y(a,M(a))$を意味する．以上のもとで，曝露の総合効果(total effect; TE)は，研究対象集団全体が曝露を受けた場合($A = 1$)と受けなかった場合($A = 0$)の比較として以下で定義される．

$$\mathrm{TE} = E[Y(1) - Y(0)] = E[Y(1, M(1)) - Y(0, M(0))]$$

総合効果は，自然な直接効果(natural direct effect; NDE)と自然な間接効果(natural

indirect effect; NIE)に分解される．自然な直接効果は以下で定義される．

$$\mathrm{NDE} = E[Y(1, M(0)) - Y(0, M(0))]$$

同様に，自然な間接効果は以下で定義される．

$$\mathrm{NIE} = E[Y(1, M(1)) - Y(1, M(0))]$$

この時，以下のような総合効果の分解が成立する．

$$\mathrm{TE} = \mathrm{NIE} + \mathrm{NDE}$$

効果の分解は個人レベルでも成立する．すなわち，対象者 i $(i = 1, \ldots, n)$に対して総合効果, 自然な直接効果・間接効果はそれぞれ $\mathrm{TE}_i = Y_i(1) - Y_i(0)$，$\mathrm{NDE}_i = Y_i(1, M_i(0)) - Y_i(0, M_i(0))$，$\mathrm{NIE}_i = Y_i(1, M_i(1)) - Y_i(1, M_i(0))$ で定義され，$\mathrm{TE}_i = \mathrm{NIE}_i + \mathrm{NDE}_i$ が成立する．

自然な直接効果と間接効果の識別に対する 1 つの十分条件は，以下の 4 つの仮定が成立することである．

$$M(a) \coprod A \mid C, \quad \forall m$$
$$Y(a, m) \coprod A \mid C, \quad \forall a, m$$
$$Y(a, m) \coprod M \mid A, C \quad \forall a, m$$
$$Y(a, m) \coprod M(0) \mid C \quad \forall a, m$$

これらの下で、自然な直接効果および自然な間接効果は以下のように表現することができる．

$$\mathrm{NDE} = \sum_c \sum_m \{E[Y \mid A = 1, M = m, c] - E[Y \mid A = 0, M = m, c]\} \times \Pr[M = m \mid A = 0, c] p(c)$$

$$\mathrm{NIE} = \sum_c \sum_m E[Y \mid A = 1, M = m, c] \times \{\Pr[M = m \mid A = 1, c] - \Pr[M = m \mid A = 0, c]\} p(c)$$

発表の後半では識別のための未測定の交絡がないという仮定が崩れた場合の感度解析方法を紹介した．紹介した方法の適用事例として，全米保健医療統計センターが公表している米国の出生証明書および乳児死亡に関するデータ解析結果を報告した．

二項反応一般化線型モデルにおける変換適合度検定統計量

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

我々は，先の研究（Taneichi et al. [5]）において，ロジスティック回帰モデルにおけるデビアンス（対数尤度比統計量）D の帰無仮説のもとでの分布に対する漸近展開式を導出し，その連続項に基づき D にバートレット修正を施した変換統計量 $\tilde{D}$ を構築した．さらにその上で，$\tilde{D}$ の検出力は D とほぼ同じで, 極限カイ二乗分布への収束が D よりも速いことを数値的に実証した．本報告では，Taneichi et al. [5] の研究内容を，ロジスティックモデルから，より一般の一般化線型モデルへ, また検定統計量もデビアンスから ϕ-ダイバージェンス (Pardo and Pardo [4]) に基づく検定統計量へと拡張をおこなった．これにより，二項反応の多様な一般化線型モデルに対する適合度検定において，標本数があまり大きくない場合であっても，極限カイ二乗分布を用いた近似検定によって，適切な検定結果を導きやすい新たな検定統計量を提案した．

2　二項反応の一般化線型モデル

一般化線型モデル (Nelder and Wedderburn [3]) を 2 項分布 $B(n,\pi)$ について考える．N 個の異なるサブグループにおける反応数に対応した確率変数 Y_α, $(\alpha=1,\ldots,N)$ が互いに独立に二項分布 $B(n_\alpha,\pi_\alpha)$, $(\alpha=1,\ldots,N)$ に従うとし，その連結関数として，単調かつ微分可能な関数 $g(\cdot)$ を用いると，二項データに対する一般化線型モデル

$$g(\pi_\alpha)=\boldsymbol{x}'_\alpha\boldsymbol{\beta},\quad(\alpha=1,\ldots,N) \tag{1}$$

が得られる．ただし, $\boldsymbol{x}_\alpha=(x_{\alpha1},\ldots,x_{\alpha p})'$, $(\alpha=1,\ldots,N)$ は共変量ベクトル, $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)'$ は未知のパラメータベクトルであり，$p<N$ とする．関数 g として特に，$g(t)=\log\{t/(1-t)\}$, $g(t)=\Phi^{-1}(t)$, $g(t)=\log\{-\log(1-t)\}$ を用いたときのモデル (1) はそれぞれ，ロジスティックモデル，プロビットモデル，補対数対数モデルとなる．ここで，$\Phi(\cdot)$ は標準正規分布の累積分布関数である．本報告で扱うモデルは，これら種々のモデルを含む一般の一般化線型モデル (1) を対象とする．

3　一般化線型モデルの適合度検定における ϕ-ダイバージェンス統計量

一般化線型モデルが正しいという帰無仮説 $H_0^g:\pi_\alpha=g^{-1}(\boldsymbol{x}'_\alpha\beta)$, $(\alpha=1,\ldots,N)$ を検定するための ϕ-ダイバージェンス検定統計量は,

$$C_\phi=2\sum_{\alpha=1}^{N}n_\alpha\left\{\hat{\pi}_\alpha^g\phi\left(\frac{\frac{Y_\alpha}{n_\alpha}}{\hat{\pi}_\alpha}\right)+(1-\hat{\pi}_\alpha^g)\phi\left(\frac{1-\frac{Y_\alpha}{n_\alpha}}{1-\hat{\pi}_\alpha}\right)\right\},$$

ただし, $\phi(\cdot)$ は $\phi(1)=\phi'(1)=0$ および $\phi''(1)=1$ を満たす $(0,\infty)$ 上での実凸関数であり, また，$\hat{\pi}_\alpha^g=\pi_\alpha(\hat{\boldsymbol{\beta}}^g)$, $(\alpha=1,\ldots,N)$ であり，$\hat{\boldsymbol{\beta}}^g=(\hat{\beta}_1^g,\ldots,\hat{\beta}_p^g)'$ は帰無仮説 H_0^g のもとでの $\boldsymbol{\beta}$ の最尤推定量である．ここで, 検定統計量 C_ϕ は, 連結関数 g に応じて異なる統計量であることを注意しておく．$n=\sum_{\alpha=1}^{N}n_\alpha$ とおくとき, 条件

$$n_\alpha/n\to\mu_\alpha,\ (\alpha=1,\ldots,N)\quad\text{as}\ \ n\to\infty, \tag{2}$$

ただし，$0 < \mu_\alpha < 1, (\alpha = 1, \ldots, N)$, $\sum_{\alpha=1}^{N} \mu_\alpha = 1$ が成り立つならば, ϕ-ダイバージェンス統計量 C_ϕ は, 帰無仮説 H_0^g のもとで $n \to \infty$ に伴って漸近的に自由度 $N-p$ のカイ二乗分布に従う．このことを用いて, 二項データが, (1) で与えられる一般化線型モデルに従うかどうかの適合度検定をおこなうことができる．

4　ϕ-ダイバージェンス統計量の改良

本報告では，C_ϕ よりもさらに極限カイ二乗分布への収束の速い統計量を構築するために，帰無仮説 H_0^g のもとでの分布の漸近展開に基づく近似として，Yarnold [6] の考え方に従い，$\Pr\{C_\phi \le x | H_0^g\} \approx J_1^{g,\phi}(x) + J_2^{g,\phi}(x)$ という近似を考える．ここで，$J_1^{g,\phi}(x)$ は連続分布を仮定した多変量エッジワース展開の項であり，$J_2^{g,\phi}(x)$ は不連続性を考慮した離散項である．C_ϕ の分布の漸近展開式を求めるために，条件 (2) の代わりに，次の仮定 1 を考える．

仮定 1 :$n_\alpha/n = \mu_\alpha$, $(\alpha = 1, \ldots, N)$ という条件を満たしながら，$n_\alpha \to \infty$, $(\alpha = 1, \ldots, N)$　as $n \to \infty$, ただし，$0 < \mu_\alpha < 1, (\alpha = 1, \ldots, N), \sum_{\alpha=1}^{N} \mu_\alpha = 1$.

すると，g^{-1} が 4 回連続微分可能, ϕ が 5 回連続微分可能であるとき，仮定 1 のもとで $J_1^{g,\phi}(x)$ は,

$$J_1^{g,\phi}(x) = \Pr\{\chi^2_{N-p} \le x\} + \frac{1}{n}\sum_{j=0}^{3} w_j^{g,\phi} \Pr\{\chi^2_{N-p+2j} \le x\} + O(n^{-2}) \tag{3}$$

という形式で評価される．ただし,χ^2_f は自由度 f のカイ二乗分布に従う確率変数を表す．また，$\sum_{j=0}^{3} w_j^{g,\phi} = 0$ という関係が成り立つ．離散項 $J_2^{g,\phi}(x)$ の評価式は非常に複雑であること，　および $J_2^{g,\phi}(x) = O(n^{-1/2})$ であることから，C_ϕ の分布に対して，(3) 式で与えられる $J_1^{g,\phi}(x)$ の近似式のみを用いて，小標本におけるカイ二乗近似の改良を考える．(3) 式に, バートレット修正および改良変換の構築と漸近展開式との関係の理論 (e.g. Fujikoshi [2]) を適用する．(3) 式において,$\phi'''(1) = -1$ かつ $\phi^{(4)}(1) = 2$ が成り立つ場合には $w_1^{g,\phi} = -w_0^{g,\phi}$ かつ $w_2^{g,\phi} = w_3^{g,\phi} = 0$ が成り立つ. このことは, バートレット修正が可能なことを意味する．よって, この場合には, C_ϕ にバートレット修正を施した変換統計量

$$C_\phi^{B*} = [1 + 2w_0^{g,\phi}\{n(N-p)\}^{-1}]C_\phi$$

を構築し, その他の場合には C_ϕ の対数形の改良変換統計量 C_ϕ^{I*} (e.g. Fujikoshi [2]) を構築した. 我々は, $w_j^{g,\phi}$, $(j = 0, 1, 2, 3)$ として, その中に含まれるパラメータ $\boldsymbol{\beta}$ に最尤推定値 $\hat{\boldsymbol{\beta}}^g$ を代入して得られる推定値 $\hat{w}_j^{g,\phi}$, $(j = 0, 1, 2, 3)$ を用いた変換統計量 $\tilde{C}_\phi^B$ および $\tilde{C}_\phi^I$ を統計量として提案した. 具体的な統計量としてパワーダイバージェンス統計量の族 R^a(Cressie and Read [1]) を考え, 変換統計量と元の統計量とのカイ二乗分布への収束の速さおよび検出力を数値的に比較した．

参考文献

[1] Cressie, N. and Read, T. R. C.: *J. R. Statist. Soc.* B, **46** (1984), 440–464.

[2] Fujikoshi, Y.: *J. Mult. Anal.*, **72** (2000), 249–263.

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

[4] Pardo, J. A. and Pardo, M. C.: *Methodol. Comput. Appl. Probab.*, **10** (2008), 357–379.

[5] Taneichi, N., Sekiya, Y. and Toyama, J.: *J. Mult. Anal.*, **102** (2011), 1263–1279.

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

二項回帰モデルの不均衡極限と変形指数型分布族

清 智也 （慶應義塾大・理工）

概要

ロジスティック回帰モデルは，応答変数が不均衡であるという設定の下で，指数型分布族に収束することが知られている．本稿では，このような現象が，ほかの二項回帰モデルに対しても普遍的に成り立つことを示す．証明は極値理論に基づいている．ロジット，プロビット，complementary log-log などのリンク関数については，収束先は指数型分布族となるが，一般には変形指数型分布族と呼ばれる分布族が現れる．

キーワード：二項回帰，極値理論，不均衡データ，ポアソン点過程．

1　はじめに：不均衡データとは？

本稿では，p 個の説明変数と 1 個の二値応答変数からなる，サイズ m の i.i.d. データ

$$\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{m}, \quad (X_i, Y_i) \in \mathbb{R}^p \times \{0, 1\},$$

を考える．ただし m, p は自然数である．$Y_i = 1$ となる i を正例，$Y_i = 0$ となる i を負例と呼ぶ．そして，$\mathcal{D}$ にもとづく条件付き確率 $P(Y_i|X_i)$ の推定を考える．

我々の興味は，データが非常に不均衡（imbalanced）な場合である．これは正例（あるいは負例）の個数がほとんどゼロという意味であり，事例としては，不正検出，医療診断，政策解析などがある (Bolton and Hand, 2002; Chawla et al., 2004; Jin et al., 2005; King and Zeng, 2001)．

本稿ではより限定的に，次のように不均衡性を定義する．

定義 1. $m \to \infty$ のもとで $\sum_{i=1}^{m} Y_i = O_P(1)$ となるとき，データ $\mathcal{D}$（の分布）は**不均衡**であるという．

i.i.d. という仮定から，データが不均衡になるためには，(X_i, Y_i) の分布がサンプルサイズ m に依存する必要がある[1]．いわゆるポアソンの少数の法則から，ある $\lambda > 0$

[1] したがって，正確には triangular array $\{(X_{m,i}, Y_{m,i})\}$ を考えていることになる．

に対して

$$P(Y_i = 1) = \frac{\lambda}{m} + o(m^{-1}), \quad m \to \infty, \tag{1}$$

が成り立つならば，$\sum_{i=1}^{m} Y_i$ は平均 λ のポアソン分布に分布収束することが分かる．さらに，次の補題が成り立つ．

補題 1. 定義 1 の意味で不均衡になるための必要十分条件は $P(Y_i = 1) = O(m^{-1})$ となることである．

なお，分布がサンプルサイズ m に依存するという設定は，それほど不自然ではない．例えば，検定論では局所対立仮説を考慮することがある．

以下では，式 (1) を満たすモデルを考察する．そのようなモデルはもちろん無数に考えられるが，ここでは次の 2 つの設定に焦点を絞る．

設定 (A) $X_i|Y_i$ の条件付き分布が m に依存しない．(Y_i の周辺分布が m に依存する．)

設定 (B) X_i の周辺分布が m に依存しない．($Y_i|X_i$ の条件付き分布が m に依存する．)

設定 (A) は，Y_i が原因，X_i が結果であるような現象を想定すると自然な仮定であると言える．一方で，回帰モデルの極限を考えるときには設定 (B) の方が直接的である．実は，不均衡性の仮定から，設定 (A) と (B) は漸近的には同等となる（4 節）．

Owen (2007) は，設定 (A) の下でロジスティック回帰を適用した場合，その最尤推定量の極限は指数型分布族の最尤推定量に収束することを示した．また，Warton and Shepherd (2010) は，モデリングの目的が本稿とは異なるものの，本質的には設定 (B) の下で，ロジスティック回帰の極限がポアソン点過程となることを指摘し，その強度が指数型分布族になることを示している．この結果については 2 節で簡単に紹介する．

本研究では，設定 (B) の下で，ロジスティック回帰以外の二項回帰モデルの極限もポアソン点過程となること，またその強度は一般に変形指数型分布族になることを示した (Sei, 2013)．その結果を報告する（3 節）．なお，漸近的な結果における正則条件は省略する．

2　ロジスティック回帰モデルの不均衡極限

ロジスティック回帰モデルは次の式で定義される：

$$P(Y_i = 1 \mid X_i = x, a, b) = \frac{e^{a+b'x}}{1 + e^{a+b'x}} \tag{2}$$

ここで，$(\alpha, \beta) \in \mathbb{R} \times \mathbb{R}^p$ を固定し，

$$a = a_m = -\log m + \alpha, \quad b = b_m = \beta$$

とおくと，

$$\begin{aligned} P(Y_i = 1 \mid X_i = x, a_m, b_m) &= \frac{\frac{1}{m} e^{\alpha + \beta' x}}{1 + \frac{1}{m} e^{\alpha + \beta' x}} \\ &= \frac{1}{m} e^{\alpha + \beta' x} + o(m^{-1}) \end{aligned}$$

となる．1 節で述べた設定 (B) の下では，X_i の周辺分布は m によらず $F(dx)$ とおける．よってベイズの定理から，

$$\begin{aligned} P(X_i \in dx \mid Y_i = 1) &= \frac{P(Y_i = 1 \mid X_i = x) F(dx)}{\int_{\mathbb{R}^p} P(Y_i = 1 \mid X_i = x) F(dx)} \\ &= \frac{e^{\alpha + \beta' x} F(dx)}{\int_{\mathbb{R}^p} e^{\alpha + \beta' x} F(dx)} + o(1) \end{aligned}$$

が成り立つ．これは $X_i | Y_i$ の条件付き分布が，自然パラメータ β の指数型分布族になることを示している．また，Y_i の周辺分布は，全確率の定理から

$$P(Y_i = 1) = \frac{1}{m} \int e^{\alpha + \beta' x} F(dx) + o(m^{-1})$$

となり，$\lambda = \int e^{\alpha + \beta' x} F(dx)$ とおくことにより式 (1) が満たされる．以上の結果を組み合わせると，ロジスティック回帰モデルは強度 $e^{\alpha + \beta' x} F(dx)$ のポアソン点過程モデルに収束することが示される (Warton and Shepherd, 2010)．

3　一般の二項回帰モデルの不均衡極限

二項回帰モデルは次の式で定義される：

$$P(Y_i = 1 \mid X_i = x, a, b) = G(a + b' x), \quad a \in \mathbb{R}, \quad b \in \mathbb{R}^p, \tag{3}$$

ここで $G(\cdot)$ は 1 次元の累積分布関数であり，その逆関数 $G^{-1}(p) = \sup\{z : G(z) \leq p\}$ はリンク関数である．分布関数 G がロジスティック分布 $G(x) = e^x/(1 + e^x)$ の場合，式 (3) は式 (2) に一致する．その他の G としては標準正規分布，(負の) ガンベル分布がよく用いられ，それぞれプロビット，complementary log-log リンク関数に対応する．以上の 3 つの例については，式 (3) の対数尤度関数は凹関数になることが知られている (Wedderburn, 1976)．

さて，唐突ではあるが，実数 q に対して，q-指数関数を

$$\exp_q(z) = \begin{cases} e^z, & \text{if } q = 1, \\ [1 + (1-q)z]_+^{1/(1-q)}, & \text{if } q \neq 1, \end{cases}$$

と定義する．ただし $[z]_+ = \max(z, 0)$，$[0]_+^{-1} = \infty$ とする．なお，q-指数関数は一般化パレート分布と同じものである．

極値理論では，次の補題が成り立つことがよく知られている．

補題 2 (de Haan and Ferreira (2006, Theorem 1.1.2 and 1.1.3))．ある数列 $c_m \in \mathbb{R}$，$d_m > 0$ および非自明な関数 $g(z)$ が存在して

$$G(c_m + d_m z) = \frac{1}{m} g(z) + o(m^{-1}), \quad m \to \infty,$$

が成り立つならば，$g(z)$ は（必要に応じて c_m, d_m を取り替えることにより）ある $q \in \mathbb{R}$ に対する q-指数関数に限られる．

よって，関数 G に対する次の仮定は自然である．

仮定 1. ある $q > 0$，$c_m \in \mathbb{R}$，$d_m > 0$ が存在して

$$G(c_m + d_m z) = \frac{1}{m} \exp_q(z) + o(m^{-1}), \quad m \to \infty,$$

が成り立つ．

例えば G がロジスティック分布の場合は $c_m = -\log m$，$d_m = 1$，$q = 1$ であり，コーシー分布の場合は $c_m = -m/\pi$，$d_m = m/\pi$，$q = 2$ である．

定理 1. $(\alpha, \beta) \in \mathbb{R} \times \mathbb{R}^p$ を固定し，仮定 1 の c_m, d_m を用いて，

$$a_m(\alpha) = c_m + d_m \alpha, \quad b_m(\beta) = d_m \beta$$

とおく．また，X_i の周辺分布は m によらず $F(dx)$ であるとする．このとき

$$\lim_{m \to \infty} P(X_i \in dx \mid Y_i = 1) = \frac{\exp_q(\alpha + \beta' x) F(dx)}{\int_{\mathbb{R}^p} \exp_q(\alpha + \beta' x) F(dx)} \tag{4}$$

が成り立つ．この右辺は q-指数型分布族（変形指数型分布族，α-分布族）と呼ばれる．

Proof. 仮定より

$$\begin{aligned} P(Y_i = 1 \mid X_i = x, a_m(\alpha), b_m(\beta)) &= G(a_m(\alpha) + b_m(\beta)' x) \\ &= G(c_m + d_m(\alpha + \beta' x)) \\ &= \frac{1}{m} \exp_q(\alpha + \beta' X_i) + o(m^{-1}). \end{aligned}$$

が成り立つ．X_i の周辺分布が $F(dx)$ のとき，ベイズの定理から

$$P(X_i \in dx \mid Y_i = 1) = \frac{\exp_q(\alpha + \beta' x)F(dx)}{\int_{\mathbb{R}^p} \exp_q(\alpha + \beta' x)F(dx)} + o(1)$$

となる． □

ロジスティック回帰のときと同様，二項回帰モデルはポアソン点過程に収束し，その強度関数は $\lambda(dx) = \exp_q(\alpha + \beta' x)F(dx)$ となる (Sei, 2013)．

4 二つの設定の漸近同等性

ここでは設定 (A), (B) が漸近的に同等であることを述べる．設定 (A) は，適当な λ, F_0, F_1 を用いて

$$\begin{aligned}
&P(Y_i = 1) = \frac{\lambda}{m} + o(m^{-1}),\\
&P(X_i \in dx \mid Y_i = 0) = F_0(dx) + o(1),\\
&P(X_i \in dx \mid Y_i = 1) = F_1(dx) + o(1)
\end{aligned}$$

と表される．ただし $o(1)$ の部分は 0 というのが設定 (A) だが，少し仮定を緩めた．

一方，設定 (B) は，適当な h, F を用いて

$$\begin{aligned}
&P(Y_i = 1 \mid X_i = x) = \frac{h(x)}{m} + o(m^{-1}),\\
&P(X_i \in dx) = F(dx) + o(1)
\end{aligned}$$

と表される．設定 (A) のときと同様，$o(1)$ の部分は少し仮定を緩めた．このように変更しても 2 節と 3 節の結果は保持される．

このとき次の定理が成り立つ．

定理 2. λ, F_0, F_1 と h, F は次の関係を持つ：

$$\begin{aligned}
&F(dx) = F_0(dx), \quad h(x) = \lambda \frac{F_1(dx)}{F_0(dx)},\\
&\lambda = \int h(x)F(dx), \quad F_0(dx) = F(dx), \quad F_1(dx) = \frac{h(x)F(dx)}{\int h(x)F(dx)}.
\end{aligned}$$

Proof. 全確率の定理とベイズの定理に基づく．設定 (A) の下では

$$\begin{aligned}
&P(Y_i = 1 \mid X_i = x) = \frac{\lambda}{m}\frac{F_1(dx)}{F_0(dx)} + o(m^{-1}),\\
&P(X_i \in dx) = F_0(dx) + o(1)
\end{aligned}$$

となる．同様に，設定 (B) の下では

$$
\begin{aligned}
P(Y_i = 1) &= \frac{\int h(x)F(dx)}{m} + o(m^{-1}), \\
P(X_i \in dx \mid Y_i = 1) &= \frac{h(x)F(dx)}{\int h(x)F(dx)} + o(1), \\
P(X_i \in dx \mid Y_i = 0) &= F(dx) + o(1)
\end{aligned}
$$

となる． □

参考文献

Bolton, R.J., Hand, D.J., 2002. Statistical fraud detection: a review. Statist. Sci. 17, 235–249.

Chawla, N.V., Japkowicz, N., Koltz, A., 2004. Editorial: special issue on learning from imbalanced data sets. ACM SIGKDD Explorations Newsletter 6, 1–6.

de Haan, L., Ferreira, A., 2006. Extreme value theory, an introduction. New York: Springer.

Jin, Y., Rejesus, R.M., Little, B.B., 2005. Binary choice models for rare events data: a crop insurance fraud application. Applied Economics 37, 841–848.

King, G., Zeng, L., 2001. Logistic regression in rare events data. Political Analisis 9, 137–163.

Owen, A.B., 2007. Infinitely imbalanced logistic regression. J. Mach. Learn. Res. 8, 761–773.

Sei, T., 2013. Infinitely imbalanced binomial regression and deformed exponential families. Preprint, arXiv:1303.7297 .

Warton, D.I., Shepherd, L.C., 2010. Poisson point process models solve the "pseudo-absence problem" for presence only data in ecology. Ann. Applied Statist. 4, 1383–1402.

Wedderburn, R.W.M., 1976. On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. Biometrika 63, 27–32.

2値判別分析におけるモデルと推定の関係について

江口 真透 （統計数理研究所 数理・推論研究系）

2 値判別の問題はさまざまな応用を含むので解析の目的も多様となる．従って統計方法もフィッシャーの線形判別とロジスティック回帰の標準的な方法から尤度に基づかないアダブースト，ランダムフォレスト，サポートベクターマシンなど多種の方法が開発されている．この発表ではそれらの統計方法をベイズリスク一致性の観点から統一的な考察を行った．

確率変数 $(\boldsymbol{X}, Y)$ を考える．ここで $\boldsymbol{X}$ は $\mathbb{R}^p$ 上の値を取り，Y は 2 値 $0, 1$ を取るとする．2 値判別問題の内容において $\boldsymbol{X}$ は特徴変数または予測変数，Y はクラスラベルと呼ばれ，目的は $\boldsymbol{X}$ に基づく Y の予測となる．同時密度関数を $p(\boldsymbol{x}, y)$ と書くとき，クラスラベル $Y = y$ の特徴変数 $\boldsymbol{X} = \boldsymbol{x}$ の条件付き確率を $P(Y = y|\boldsymbol{x})$，特徴変数 $\boldsymbol{X}$ のクラスラベル $Y = y$ の条件付き密度関数を $p(\boldsymbol{x}|Y = y)$ とすると，対数尤度比 $\Lambda(\boldsymbol{x}) = \log p(\boldsymbol{x}, 1)/p(\boldsymbol{x}, 0)$ は

$$\Lambda(\boldsymbol{x}) = \log \frac{P(Y = 1|\boldsymbol{x})}{P(Y = 0|\boldsymbol{x})} \quad \text{または} \quad \Lambda(\boldsymbol{x}) = \log \frac{p(\boldsymbol{x}|Y = 1)P(Y = 1)}{p(\boldsymbol{x}|Y = 0)P(Y = 0)} \tag{1}$$

と書ける．識別子 $f : \boldsymbol{X} \mapsto Y$ は判別関数 $F(\boldsymbol{X})$ と閾値 c から $f(\boldsymbol{X}) = \mathrm{H}(F(\boldsymbol{X}) - c)$ と構成される．ここで $\mathrm{H}(\cdot)$ はヘビサイド関数を表す．このように識別子は直接に構成されるのではなく判別関数によって構成される．

特徴量 $\boldsymbol{X}$ に基づく Y の予測は (1) の Λ の 2 つの表現に基づき，前者は予測的アプローチ，後者は診断的アプローチを導く．予測的アプローチはロジスティック回帰モデルのように回帰関数 $P(Y = y|\boldsymbol{x})$ を直接モデル化によって回帰関数の推定問題として扱い，アダブーストやサポートベクターマシンなどはこれに含まれる [1 – 5]．一方で診断的アプローチはフィッシャー線形判別のようにクラスラベル条件付き密度関数 $p(\boldsymbol{x}|Y = y)$ を推定する問題として扱い，2 標本検定と密接な関係がある．この文脈では帰無仮説はクラスラベル条件付き分布の同等性 $H : p(\boldsymbol{x}|Y = 1) = p(\boldsymbol{x}|Y = 0)$ となる．2 標本 t-検定，ウィルコクソン順位和検定が関係してくる [4]．特にフィッシャー線形判別は本質的には，t-統計量の最大化によって得られることを考察した．

密度関数 $p(\boldsymbol{x}, y)$ からランダムサンプル $(\boldsymbol{X}_i, Y_i)_{i=1}^n$ が得られたとせよ．このとき，ロジスティックモデル

$$P(Y = y|\boldsymbol{x}, \boldsymbol{\beta}) = \frac{e^{y(\boldsymbol{\beta}_1^T \boldsymbol{x} + \beta_0)}}{1 + e^{\boldsymbol{\beta}_1^T \boldsymbol{x} + \beta_0}} \tag{2}$$

が仮定されたとしよう．条件付き対数尤度関数

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log \frac{e^{Y_i(\boldsymbol{\beta}_1^T \boldsymbol{X}_i + \beta_0)}}{1 + e^{\boldsymbol{\beta}_1^T \boldsymbol{X}_i + \beta_0}} \tag{3}$$

に基づく解析が広く使われている．一方で尤度に基づかない方法も近年，活発な進展がみられる．その典型として指数ロス関数

$$L_{\exp}(F) = \frac{1}{n} \sum_{i=1}^{n} \exp\{(1 - 2Y_i)F(\boldsymbol{X}_i)\} \tag{4}$$

を考えよう．このロス関数はアダブースト・アルゴリズムを導入するために考えられたものである [2, 3]．識別子の集合 $\mathcal{C}$ を予め用意して t-ステップでの判別関数 F_t から次のステップの判別関数を $F_{t+1}(\boldsymbol{x}) = F_t(\boldsymbol{x}) + \alpha_t f_t(\boldsymbol{x})$ と更新する．ここで指数ロス関数を使って

$$(\alpha_t, f_t) = \underset{(\alpha, f) \in \mathbb{R} \times \mathcal{C}}{\operatorname{argmin}} \; L_{\exp}(F_t + \alpha f) \tag{5}$$

と定める．上の (5) の最小化は初等的な解があることが示され，このステップワイズな反復より最終形は $\hat{F}(\boldsymbol{x}) = \sum_{t=1}^{T} \alpha_t f_t(\boldsymbol{x})$ が得られる．識別子の集合 $\mathcal{C}$ が判別問題に対してうまく選択されたならば強力な判別解析が得られることが知られている．期待指数ロス関数は次のように表される．

$$\mathbb{E}\{L_{\exp}(F)\} = \int \big[\exp\{-F(\boldsymbol{x})\}p(Y=1|\boldsymbol{x}) + \exp\{F(\boldsymbol{x})\}p(Y=0|\boldsymbol{x})\big]p(\boldsymbol{x})d\boldsymbol{x}$$

命題 1. すべての判別関数の空間を $\mathcal{F}$ と表す．このとき，$F^*(\boldsymbol{X}) = \frac{1}{2}\Lambda(\boldsymbol{X})$ とすると次が成立する．

$$\mathbb{E}\{L_{\exp}(F^*)\} = \min_{F\in\mathcal{F}} \mathbb{E}\{L_{\exp}(F)\}. \tag{6}$$

注意 1. 線形判別関数 $F^*(\boldsymbol{x}) = \frac{1}{2}(\beta_1^T\boldsymbol{x} + \beta_0)$ を考えると (1) の 1 番目の尤度比 Λ の表現から

$$P(Y=y|\boldsymbol{x}) = \frac{e^{y(\beta_1^T\boldsymbol{x}+\beta_0)}}{1+e^{\beta_1^T\boldsymbol{x}+\beta_0}} \tag{7}$$

となりロジスティック回帰モデルに帰着される [1].

90 年代より機械学習の分野で活発に考察されたアダブーストとサポートベクターマシンは必ずしも尤度に基づかない方法により拡大されている，一方で統計学においては 90 年代より回帰分析の内容で混合効果モデルの研究が盛んに進められてきた．ラプラス近似法や MCMC 法によるプログラムによる普及から医学統計，生物統計，生態学において混合効果モデルの解析が標準的なものとして広く浸透してきている．不思議なことに，この機械学習の流れと統計学の流れは互いに接点を持たない．その理由の一つとして考えられることは周辺尤度の考えを他のロス関数に拡大することの困難さが挙げられる．これについて以下のよう予測アプローチ上で考察しよう．判別関数 F に対してあるロス関数 $L(F)$ がベイズリスク一致性を満たすとし，線形判別モデル $F_{\boldsymbol{\beta},\boldsymbol{b}}(\boldsymbol{x},\boldsymbol{z}) = \boldsymbol{\beta}^T\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{z} + \alpha$ を仮定すると連想モデルは

$$P(Y=y|\boldsymbol{x},\boldsymbol{z}) = \frac{e^{y(\boldsymbol{\beta}^T\boldsymbol{x}+\boldsymbol{b}^T\boldsymbol{z}+\alpha)}}{1+e^{\boldsymbol{\beta}^T\boldsymbol{x}+\boldsymbol{b}^T\boldsymbol{z}+\alpha}} \tag{8}$$

で与えられる．ここで $\boldsymbol{\beta}^T\boldsymbol{x}$ を固定効果，$\boldsymbol{b}^T\boldsymbol{z}$ をランダム効果とする．この設定の下でロス関数 L に基づく方法は未だ解決に至ってないと思われる．今後，統計学と機械学習の分野で密接な議論を通して判別分析，パタン認識の新たな方向が模索されることが期待される．

参考文献

[1] S. Eguchi and J. Copas. A class of logistic-type discriminant functions. *Biometrika* 2002, *89*, 1, 1–22.

[2] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Computer and System Sciences*, 55, 119–139, 1997.

[3] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. *Ann. Statist.* 28, 337–407, 2000.

[4] O. Komori. A boosting method for maximization of the area under the ROC curve. *Ann. Inst. Statist. Math.*, 63, 961–979, 2011

[5] V. N. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.

Nonparametric independence screening and structural identification for ultra-high dimensional longitudinal data

Ming-Yen Cheng[1], Toshio Honda[2], Jialiang Li[3] and Heng Peng[4]

[1] Department of Mathematics, National Taiwan University
[2] Graduate School of Economics, Hitotsubashi University
[3] Department of Statistics and Applied Probability, National University of Singapore
[4] Department of Mathematics, Hong Kong Baptist University

Ultra-high dimensional longitudinal data are increasingly common and the analysis is challenging both theoretically and methodologically. The purpose of this paper is to offer an automatic procedure in hunting for a sparse semivarying coefficient model, which has been widely accepted in applications.

Our idea is to first reduce the number of covariates to a reasonable order by employing a screening method, and then identify both the varying and constant coefficients using a group SCAD estimator. Based on the results, we then refine the group SCAD estimator by accounting for the within-subject covariance function, which is estimated nonparametrically. The screening procedure is based on working independence and B-spline marginal varying coefficient models. Under weaker conditions than existing ones, we show that with high probability only irrelevant variables will be screened out and the number of remaining variables can be bounded by a moderate order.

Our group SCAD regularized B-spline estimator detects the constant and varying effects simultaneously and possesses the consistency, sparsity and oracle properties. The refined semivarying coefficient model is semiparametric efficient as it employs local linear smoothing, nonparametric estimation of the covariance structure based on residuals obtained by assuming working independence and profile least squares estimation.

We also suggest ways to implement the methods and to select the tuning parameters and the smoothing parameters. In summary, the methodology and theory are new and powerful, and the methods are fully automatic and readily applicable in practice. An extensive simulation study is summarized to demonstrate its finite sample performance and the yeast cell cycle data is analyzed.

Information Criteria Based on Quasi-likelihood with Application to Over-dispersed Data

Yiping Tang, and Jinfang Wang
Graduate School of Science, Chiba University

1 Semiparametric Kullback-Leibler information

We shall consider regression analysis in the framework of generalized linear models. We will replace the full distributional assumptions by the assumptions on mean $\mu_i(\boldsymbol{\beta})$ and variance $\phi V(\mu_i(\boldsymbol{\beta}))$.Thus, the we have an infinite dimensional models space, containing all distribution functions satisfying these two moment restrictions. Extending the idea of AIC, we thus consider the following problem by projecting the true distribution function to the semiparametric model using the Kullback-Leibler information as a quasi-distance function:

$$\begin{aligned} v(\boldsymbol{\beta},\phi) = \quad & \inf_{g\in\mathcal{G}} \int \log\left(\frac{h(y)}{g(y)}\right) h(y)\ dy \\ & \text{subject to } \int m(y,\boldsymbol{\beta},\phi)g(y)\ dy = 0 \text{ and } \int g(y)\ dy = 1. \end{aligned} \tag{1.1}$$

Using the results in convex analysis (e.g., Kitamura, 2006), the solution to infinite dimensional optimization problem $v(\boldsymbol{\beta},\phi)$ can be equivalently written as the solution to the finite dimensional optimization problem $v^*(\boldsymbol{\beta},\phi)$ as follows:

$$v^*(\boldsymbol{\beta},\phi) = \sup_{\lambda\in\mathbb{R},\boldsymbol{r}\in\mathbb{R}^2} \left[1+\lambda+\int \log\left(-\lambda-\boldsymbol{r}'m\left(y,\boldsymbol{\beta},\phi\right)\right) h(y)\ dy\right], \tag{1.2}$$

which leads us to consider the following semiparametric Kullback-Leibler information

$$\mathrm{SKL}(\boldsymbol{\theta}) = \sup_{\boldsymbol{r}\in\mathbb{R}^2} \int \rho(y,\boldsymbol{\theta},\boldsymbol{r})h(y)\ dy, = \sup_{\boldsymbol{r}\in\mathbb{R}^2} \int \log\left(1+\boldsymbol{r}'m(y,\boldsymbol{\beta},\phi)\right) h(y)\ dy\,. \tag{1.3}$$

2 The semiparametric information criterion

The minimizer of $\mathrm{SKL}(\boldsymbol{\theta})$ is a solution to the following saddle point problem:

$$\boldsymbol{\theta}_* = \arg\min_{\boldsymbol{\theta}\in\Theta} \sup_{\boldsymbol{r}\in\mathbf{L}(\boldsymbol{\theta})} E_H\left[\rho(y,\boldsymbol{\theta},\boldsymbol{r})\right], \tag{2.4}$$

where Θ denotes the parameter space for $\boldsymbol{\theta}$, $\mathbf{L}(\boldsymbol{\theta}) = \{\boldsymbol{r} : \boldsymbol{r}'m(y,\boldsymbol{\theta}) \in \mathcal{I}\}$, and $\mathcal{I}$ is an open interval containing zero. To find $\boldsymbol{\theta}_*$, the expectation of $\rho(y,\boldsymbol{\theta},\boldsymbol{r})$ is first maximized for given $\boldsymbol{\theta}$, so that $\boldsymbol{\theta}_*$ satisfies $E_H\left[\rho_1\left(y,\boldsymbol{\theta},\boldsymbol{r}\right)m(y,\boldsymbol{\theta})\right] = \mathbf{0}$, where $\rho_1\left(y,\boldsymbol{\theta},\boldsymbol{r}\right)$ is the first derivative of $\rho(.)$ with respect to $\boldsymbol{r}'m(y,\boldsymbol{\theta})$. Nest, since $\boldsymbol{\theta}_*$ is the minimizer of $E_H\left[\rho(y,\boldsymbol{\theta},\boldsymbol{r})\right]$, so that $E_H\left[\rho_1\left(y,\boldsymbol{\theta},\boldsymbol{r}\right)M(y,\boldsymbol{\theta})'\boldsymbol{r}(\boldsymbol{\theta})\right] = \mathbf{0}$, is satisfies, where $M(y,\boldsymbol{\theta}) = \partial m(y,\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$.

The solutions to the above saddle point problem is not computable because it involves the expectation with respect to the true known density function. One way to avoid this

is to replace the true distribution function by an empirical distribution function. This solution can be shown to be asymptotically equivalent to $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}', \widehat{\phi})'$ with $\widehat{\boldsymbol{\beta}}$ being the quasi-likelihood estimator and $\widehat{\phi}$ the estimator from the usual Pearson residual. Since the $\widehat{\boldsymbol{\theta}}$ is also a consistent estimator, it is reasonable to use $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{r}}(\widehat{\boldsymbol{\theta}}))$ instead. The statistical problem now is to construct an estimator for the expected semiparametric information defined as follows:

$$E_H[\mathrm{SKL}(\widehat{\boldsymbol{\theta}})] = E_H\left[\int \rho(y, \widehat{\boldsymbol{\theta}}, \tilde{\boldsymbol{r}}(\widehat{\boldsymbol{\theta}}))h(y)\,dy\right]. \tag{2.5}$$

Now we are able to state our main results.

The proposed criterion SIC has the following properties.

THEOREM 2.1 *Suppose that the model is correctly specified. Let k be the dimension of $\boldsymbol{\theta}$. Then the following quantity*

$$\mathrm{SIC} = \sum_{i=1}^{n} \rho\left(y_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{r}}(\widehat{\boldsymbol{\theta}})\right) + k. \tag{2.6}$$

is an asymptotically unbiased estimator of the expected semiparametric information.

When models are possibly misspecified, we can extend the above results as follows.

THEOREM 2.2 *Let*

$$\widehat{\mathrm{S}} = \frac{1}{n}\sum_{i=1}^{n}\left[\boldsymbol{\Upsilon}_i\left(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{r}}(\widehat{\boldsymbol{\theta}})\right)\boldsymbol{\Upsilon}_i\left(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{r}}(\widehat{\boldsymbol{\theta}})\right)'\right], \quad \widehat{\mathrm{Q}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial \boldsymbol{\Upsilon}_i\left(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{r}}(\widehat{\boldsymbol{\theta}})\right)}{\partial(\boldsymbol{\theta}', \boldsymbol{r}')'}\right].$$

Then the following quantity,

$$\mathrm{SIC}_T = \sum_{i=1}^{n} \rho\left(y_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{r}}(\widehat{\boldsymbol{\theta}})\right) - \mathrm{trace}(\widehat{\mathrm{S}}\widehat{\mathrm{Q}}^{-1}), \tag{2.7}$$

under regularity conditions, is an asymptotically unbiased estimator of the expected semiparametric information.

SIC_T may be understood as a semiparametric extension of the Takeuchi information criterion (Takeuchi, 1976) for parametric likelihood analysis.

一部の観測領域でランダムな欠測のあるデータへの混合分布モデルの適用

札幌学院大学 経済学部　中村　永友
城西大学 理学部　土屋　高宏
統計数理研究所 モデリング研究系　上野　玄太

1　はじめに

欠測データに対しては，典型的なパラメトリックな統計的モデリングとして値打ち切り (censored) や切断 (truncated) を考慮した分析方法がある (中村他, 2005)．本報告では，ある特定の領域で分析上無視できない欠測データがあり，さらにその領域では正常に測定されたデータも存在している，という問題を扱う．例えば機器の調子が悪く，ある特定の測定範囲で記録されたデータもあれば，うまく記録されなかったデータもあるという状況である．これは部分的な一部の領域でランダムな欠測がある (partially missing at random) という視点でモデリングが可能である．このようなデータを分析するには，従来の値打ち切りや切断の統計モデルをそのまま適用できない．本報告の目的は，このような問題に対する統計モデルを示し，さらに基礎となる確率分布を正規混合分布に拡張することである．

2　基礎となる統計モデルと尤度関数

確率分布 $f(\cdot)$ の定義域内の任意の領域 R で全くデータが観測されなかったという統計モデルは，次のように構成することができる．領域 R を欠測領域と呼ぶこととし，R において欠測数 m を既知とするとき，観測データ x と定義関数

$$\delta = \begin{cases} 1: \ x \in \bar{R} \ \text{(測定値あり)} \\ 0: \ x \in R \ \text{(測定値なし = 欠測)} \end{cases}$$

を用いると，考え得る 1 つの統計モデルは同時密度関数

$$X \sim f(x|\theta)^{\delta} \cdot p(x|\theta, R)^{1-\delta}$$

である．ここで，

$$p(x|\theta, R) = \int_R f(x)dx$$

である．これに基づいて尤度関数を構成して，目的のパラメータを推定することができる (中村他, 2005)．

次に，領域 R で観測と欠測が同時にあることを考える．この状況の 1 つの考え方として，R のみでランダムな欠測がある，という統計モデルが構成できる．この状況は，領域 $\overline{R} = (-\infty, c)$ でデータが正常に $n_{\overline{R}}$ 個観測され，領域 $R = [c, \infty)$ では n_R 個データが観測され，m 個欠測しているとする (図 1)．

この状況から自然に導かれる尤度関数は以下の通りである．

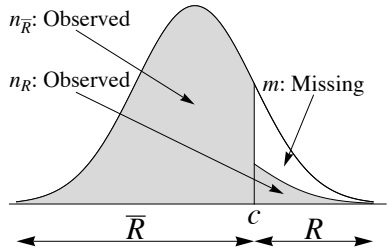

図 1: 領域 R での欠測と観測されたデータ数 $(n_{\overline{R}}, n_R, m)$.

$$
\begin{aligned}
L_0 &= \prod_{i=1}^{n_{\overline{R}}} f(x_i|\theta) \cdot \prod_{j=1}^{n_R} f(x_j|\theta) \cdot \left\{ \int_R f(t|\theta)dt \cdot \frac{m}{n_R + m} \right\}^m \\
&= \prod_{i=1}^{N} f(x_i|\theta) \cdot \left\{ \int_R f(t|\theta)dt \cdot \frac{m}{n_R + m} \right\}^m . \qquad (1)
\end{aligned}
$$

ここで，$N = n_{\overline{R}} + n_R$ は測定されたデータ数，式 (1) の中括弧内は R における欠測率を表す確率である．欠測数 m が既知の場合は定数として扱い，未知の場合はパラメータとして推定を行うことができる (中村他, 2013).

3 混合分布モデルへの拡張

基礎となる分布を r 個の成分分布を持つ混合正規分布モデル

$$f(x|\mathbf{\Theta}) = \sum_{k=1}^{r} \pi_k f_k(x|\theta_k)$$

とする．第 k 成分分布 $f_k(\cdot|\theta_k)$ は 1 次元正規分布，$\sum \pi_k = 1$, $\pi_k > 0$, $\theta_k = \{\mu_k, \sigma_k^2\}$, $\mathbf{\Theta} = \{\pi_1, \ldots, \pi_r, \theta_1, \ldots, \theta_r\}$ である．(1) 式に基づいて，混合分布の各成分分布に対して領域 R を考慮した尤度関数が構成できる．

$$L = \prod_{i=1}^{N} f(x_i|\mathbf{\Theta}) \times \prod_{k=1}^{r} \left\{ \int_R \pi_k f_k(t|\theta_k)dt \times \frac{m_k}{n_{R_k} + m_k} \right\}^{m_k} .$$

n_{R_k} と m_k は第 k 成分分布の R における観測数と欠測数である．これによる対数尤度関数から，EM 法で推定するためのパラメータの更新式を構成することができる．また，m_k は既知でも未知でも両方に対してパラメータ推定でき，さらに未知の場合は欠測数 m_k の推定も可能である．

4 数値実験

4.1 正規分布モデル

データ生成の統計モデルを標準正規分布として，欠測領域は $R = (0, \infty)$，欠測率（欠測領域内の欠測の割合）は $q = 0.9, 0.75, 0.5, 0.25, 0.1$，発生させるデータ数を 1000 として，$R$ 内で q の各値に対して数値実験を行った．実験 (1) は欠測数を既知，実験 (2) は未知として推定している．実験回数は 100 回として，その平均を示したのが表 1 である．

表 1: 正規分布モデルでの数値実験

	q	$\hat{\mu}$	$\hat{\sigma}^2$	$n_{\bar{R}}$	n_R	$\hat{m}$	$n_{\bar{R}}+n_R+\hat{m}$
実験 (1)	0.90	−0.003	0.990	500.9	49.5	450.0	1000
	0.75	0.000	1.002	498.8	126.3	375.0	1000
	0.50	−0.001	1.000	501.3	247.7	251.0	1000
	0.25	−0.005	1.003	503.5	371.8	124.7	1000
	0.10	−0.004	0.999	500.9	450.3	48.8	1000
実験 (2)	0.90	−0.010	0.996	501.8	49.1	449.0	999.9
	0.75	−0.001	1.001	500.2	124.5	377.7	1002.4
	0.50	0.005	1.004	495.9	251.8	250.3	998.0
	0.25	0.002	0.999	499.2	373.5	129.2	1001.9
	0.10	−0.001	0.998	500.6	449.2	53.9	1003.7

4.2 混合正規分布モデル (1)

データ生成の統計モデルを 2 成分からなる混合正規分布

$$\frac{1}{2}\{N(x|0,1)+N(x|3,1)\}$$

として，欠測領域は $R=(4,\infty)$，欠測率（欠測領域内の欠測の割合）は $q=0.9,0.75,0.5,0.25$，発生させるデータ数を 1000 として，R 内で q の各値に対して数値実験を行った．実験 (3) は欠測数を既知，実験 (4) は未知として推定している．実験回数は 100 回として，その平均を示したのが表 2 である．

表 2: 混合正規分布モデルでの数値実験 1

	q	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{p}$	$N+\hat{m}$	N	$\hat{m}$	m^0	$\hat{m}_1$	$\hat{m}_2$	m_1^0	m_2^0
実験 (3)	0.25	0.505	0.495	0.021	3.023	1.020	0.982	0.0195	1000	980.5	19.5	19.8	0.0091	19.5	0.004	19.8
	0.50	0.501	0.499	0.005	3.011	1.008	0.991	0.0407	1000	959.3	40.7	39.7	0.0173	40.7	0.008	39.7
	0.75	0.500	0.500	0.004	2.992	1.017	1.017	0.0598	1000	940.3	59.7	59.5	0.0317	59.7	0.012	59.5
	0.90	0.496	0.504	−0.005	2.990	0.990	1.017	0.0716	1000	928.6	71.4	71.4	0.0265	71.4	0.014	71.4
実験 (4)	0.25	0.497	0.503	−0.009	3.003	0.991	1.022	0.0233	1003.9	980.3	23.6	19.8	0.0097	23.6	0.004	19.8
	0.50	0.502	0.498	0.002	3.013	1.013	0.996	0.0404	999.9	959.2	40.7	39.7	0.0193	40.7	0.008	39.7
	0.75	0.496	0.504	−0.001	3.010	1.000	1.042	0.0637	1005.9	941.5	64.5	59.5	0.0232	64.5	0.012	59.5
	0.90	0.502	0.498	0.000	2.989	1.005	0.999	0.0690	998.4	929.0	69.5	71.4	0.0326	69.4	0.014	71.4

4.3 混合正規分布モデル (2)

データ生成の統計モデルを 2 成分からなる混合正規分布

$$\frac{1}{2}\{N(x|0,1)+N(x|6,1)\}$$

として，欠測領域は $R=(1,5)$，欠測率（欠測領域内の欠測の割合）は $q=0.25,0.5,0.75,0.9$，発生させるデータ数を 1000 として，R 内で q の各値に対して数値実験を行った．実験 (5) は欠測数を既知，実験 (6) は未知として推定している．実験回数は 100 回として，その平均を示したのが表 3 である．

表 3: 混合正規分布モデルでの数値実験 2

	q	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{p}$	$N+\hat{m}$	N	$\hat{m}$	m^0	$\hat{m}_1$	$\hat{m}_2$	m_1^0	m_2^0
実験 (5)	0.25	0.500	0.500	-0.005	5.997	1.006	0.998	0.03938	1000	960.7	39.3	39.7	19.6	19.7	19.8	19.8
	0.50	0.500	0.500	-0.005	5.994	1.011	0.999	0.08011	1000	919.6	80.4	79.3	40.0	40.4	39.7	39.7
	0.75	0.501	0.499	-0.001	6.000	1.002	1.001	0.11908	1000	881.1	118.9	119	59.5	59.4	59.5	59.5
	0.90	0.500	0.500	0.003	5.995	0.999	0.995	0.14399	1000	855.8	144.2	142.8	72.2	72.0	71.4	71.4
実験 (6)	0.25	0.500	0.500	0.001	5.996	0.995	0.997	0.04040	1001.7	960.6	41.1	39.7	20.5	20.6	19.8	19.8
	0.50	0.501	0.499	0.004	5.997	1.005	0.998	0.07917	999.6	920.0	79.7	79.3	40.1	39.6	39.7	39.7
	0.75	0.500	0.500	-0.001	5.990	1.009	0.989	0.11848	1001.4	882.2	119.3	119	59.8	59.5	59.5	59.5
	0.90	0.500	0.500	0.005	5.989	1.016	1.026	0.14814	1005.4	855.7	149.7	142.8	74.4	75.4	71.4	71.4

5　試験完成時間データへの適用

情報教育（excel）の試験完成時間データに対して，提案手法を適用する．課題が完成した時刻がファイル属性として記録されたファイルが提出され，その完成時間が分析対象データとなる．締め切り時間まで解答して，その後自らツールで採点する過程を経るため，記録される時間が締め切り時間後になることもあり得る．また，各 PC の時計が狂っていたり，多人数で受験するため，試験監督者の目を盗んで，締め切り時間後も解答し，採点ということもあり得る．このようなことから，締め切り時間後のデータも存在している．実験計画と実際のデータ取得における理想と現実のギャップが顕在している．

このデータの取得背景として，あと少し時間があれば合格する不合格者（潜在合格者）が存在し，一方，全く合格の可能性のない学生（完全不合格者）もいる．

このデータ分析の目的は，このデータから潜在合格者数を推定することである．これを行うことの仮定は次の通りである．測定データのヒストグラムの観察から，合格者は 2 つの分布が混合しているように見える．１つは実力があって自力で容易に課題を完成できる群（自力合格群）と，もう 1 つはある程度時間がかかって試行錯誤的に課題を完成できる群（試行錯誤群）とみることができる．試行錯誤群の一部は潜在合格者で，時間が足りずに不合格になっていると考えることができる．この人数の推定をすることが目的である．

データの概要は次の通りである．正規合格者は 530 人（内，時間内合格は 516 人，時間後の合格者は 14 人），不合格は 269 人である．不合格者には，潜在合格者と完全不合格者がいると考えて，この潜在合格者数を推定する．

適用結果を図 2 に示す．横軸が時間で，右方向にある縦棒が締め切り時間である．使用したデータは時間の測定値のみを用いていて，欠測数は用いていない．完全不合格者群と潜在合格者を分離する情報がないためである．さらに，2 つの正規分布に対して，分散や混合比率に対して何も条件をつけずにパラメータ推定した．パラメータの推定値は，$\hat{\pi}_1 = 0.359$, $\hat{\mu}_1 = -34.7$, $\hat{\mu}_2 = -10.2$, $\hat{\sigma}_1^2 = 62.1$, $\hat{\sigma}_2^2 = 166.6$ であった．この結果から，潜在合格者数は 81 人と推定された．

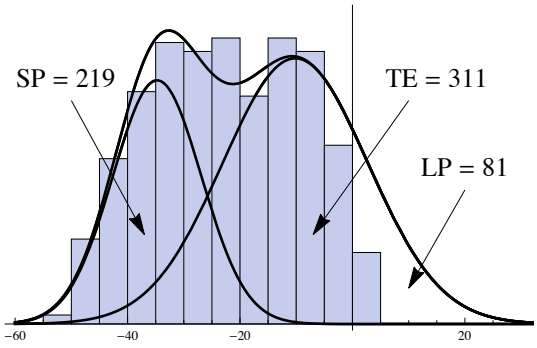

図 2: 試験完成データに対する提案モデルの当てはめ
SP: 自力合格者群，TE: 試行錯誤群, LP: 潜在合格者．

参考文献

[1] 中村永友, 金子真紀子, 于秉柯 (2009). 欠測がある基本味質データの平均・相関構造の探索, 計量生物学, **30**(1), 55-67.

[2] 中村永友, 土屋高宏, 上野玄太 (2013). 一部の観測領域でランダムな欠測のあるデータに対する混合分布モデルのあてはめ, 2013 年度統計関連学会連合大会 予稿集.

[3] 中村永友, 上野玄太, 樋口知之, 小西貞則 (2005). 欠損混合分布モデルとその応用, 応用統計学, **34**(2), 57-75.

カーネル法に基づく順位検定の連続化について

前園 宜彦　九州大学数理学研究院

1　序

$X_1, \cdots, X_n$ を $F(x-\theta)$ からの無作為標本とする．ただし F は $F(-x) = 1-F(x)$ の対称な分布で密度関数を持つとする．このとき帰無仮説 $H_0 : \theta = 0$ vs. 対立仮説 $H_1 : \theta > 0$ の一標本検定問題を考える．ノンパラメトリックな検定としては，符号検定 $S = \#\{X_i \geq 0; i = 1, 2, \cdots, n\}$ やウィルコクソンの符号付き順位検定 $W = \#\{X_i + X_j \geq 0; 1 \leq i \leq j \leq n\}$ が良く利用されている．検定は実現値 $s,\ w$ に対して有意確率 $P_0(S \geq s),\ P_0(W \geq w)$ を求めて，その値が小さい時に帰無仮説 H_0 を棄却することになる．しかしいくつかの文献（Lehmann (1975), Brown et al.(2001) 等）で指摘されているように $S,\ W$ は離散型の分布を持つために，標本数が小さい時は多くの場合 W を利用する方が S よりも有意確率は小さくなる．正規近似の半数補正を使って有意確率を求めると，この問題は少し改善される．

分布の離散性を克服するために，Brown et al. (2001) は平滑化メディアンを提案し，その中で平滑化符号検定も議論している．また母集団分布が正規分布のときは，Pitman の漸近相対効率の意味で符号検定より優れていることを示している．平滑化の影響で，検定統計量は distribution-free ではなくなるために，近似分布及び Edgeworth 展開を求めている．

他方前園&魯 (2013) は，分布関数のカーネル型推定量を用いて，連続化符号検定を提案し，有意確率の近似を議論している．本報告ではウィルコクソンの符号付き順位検定のカーネル関数を使った連続化を提案し，その有意確率の近似について講演する．

2　ウィルコクソンの符号付き順位検定の連続化

本講演ではカーネル型統計量に基づいて，ウィルコクソンの符号付き順位と Pitman の効率の意味で同等な検定を議論する．前園&魯 (2013) は符号検定の連続化として

$$\tilde{S} = \frac{1}{n}\sum_{i=1}^{n} K\Big(-\frac{X_i}{h_n}\Big) - \frac{1}{2}$$

を提案している．ただし $k(u)$ は 4 次のカーネルとするとき

$$K(t) = \int_{-\infty}^{t} k(u)du$$

である．また h_n はバンド幅で，$h_n \to 0,\ nh_n \to \infty$ である．帰無仮説 H_0 が正しい時 $\tilde{S}$ は 0 に近い値をとる確率が大で，対立仮説が正しい時は負の値をとる確率が大になる．したがって検定は，実現値 $\tilde{s}$ に対して有意確率

$$P_0(\tilde{S} \leq \tilde{s})$$

の大きさを検討することになる．

ウィルコクソンの符号付き順位の連続化として

$$\tilde{W} = \sum_{1 \leq i \leq j \leq n} K\Big(-\frac{X_i + Xj}{h_n}\Big)$$

を考える．対立仮説が正しい時には小さい値をとる確率が大きくなるので，実現値 $\tilde{w}$ に対して

$$P_0(\tilde{W} \leq \tilde{w})$$

が小さい時に H_0 を棄却することになる．

3　検定統計量の性質

検定統計量 $\tilde{W}$ は連続化したために distribution-free ではなくなるが，帰無仮説 H_0 の下での漸近分散は母集団分布 F に依存しない．また Pitman の漸近相対効率はウィルコクソンの符号付き順位検定 W と同じである．

定理 カーネル関数 $k(u)$ は 4 次のカーネル，すなわち

$$\int_{-\infty}^{\infty} k(u)du = 1 \quad , \int_{-\infty}^{\infty} k(u)u^j du = 0, (j = 2, 3), \int_{-\infty}^{\infty} k(u)u^4 du \neq 0$$

を満たすとする．また $h_n = cn^{-1/4}$ $(c > 0)$ とするとき次が成立する．
(1)

$$V_0(\tilde{W}) = \frac{n^3}{12} + O(n^2), \quad E_\theta(\tilde{W}) = \frac{n^2}{2}\int_{-\infty}^{\infty} F(x + 2\theta)f(x)dx + O(n)$$

(2)

$$\frac{\tilde{W} - E_\theta[\tilde{W}]}{\sqrt{V_\theta(\tilde{W})}} \xrightarrow{L} N(0, 1), \quad (n \to \infty)$$

また帰無仮説 H_0 の下でも，標準化した統計量の漸近分布は標準正規分布 $N(0, 1)$ である．
(3) ウィルコクソンの符号付き順位検定 W と連続化検定 $\tilde{W}$ は同じ Pitman の漸近相対効率を持つ．

$\tilde{W}$ の分散の主要項は F に依存しないので，直接正規近似を利用して有意確率の評価が可能である．$\tilde{W}$ は連続化しているので，distribution-free ではないが，漸近分布及び Edgeworth 展開の利用ができて，カーネル関数 $k(\cdot)$ をうまくとることにより正規近似を精密化も可能である．

参考文献

[1] Brown, B.M., Hall, P. and Young, G.A.(2001), Biometrika, Vol.88, 519-534.

[2] Lehmann, E.L. (1975) Nonparametric, Holden-Day.

[3] 魯 & 前園 (2013) 符号検定の平滑化と有意確率の近似について，２０１３年日本数学会春期年会