混合効果モデルと小地域統計

久保川達也
(東京大学)

[1] 地域統計と小地域推定

かつてカナダのトロントに滞在していたときに，Toronto Sun (地元の一般紙) の一面にトロント市の小地域別平均年齢の地図がカラーで掲載され，高齢化が進んでいる地域など一目瞭然にわかるものであった。
カナダ統計局には小地域統計の専門的なグループがあり，この分野の世界的リーダーである J.N.K. Rao はこうしたグループと関わりを持ち，現場と理論とを往復させながら研究を展開してきた。
我が国でも，地域別平均年齢の予測など地域統計の重要性は増している。
小地域推定の研究分野は，時代の変化に伴う現場からの要請に応じて新たな統計手法の開発とそれに伴う理論研究や計算手法の展開を行いつつ発展している。

調査区全体の特性を調べるためにとられたデータを利用して，地域ごとの特性値を推定したい。
そのとき，狭い地域や人口が粗な地域に対しては十分なデータがとられていないため，その地域だけのデータでは特性値の十分な推測ができない。
このような状況での推定問題を，小地域推定という。

[2] 集計データを利用するためのモデル

小地域推定の問題を解決する方法は，周辺地域のデータを組み込んで推定精度を高めること。
そのために利用されるのが線形混合モデル (Linear Mixed Model, LMM) である。

Fay-Herriot モデル：集計データに基づいたモデル
m：小地域の個数，$i = 1, \ldots, m$
y_i：i 番目の小地域の集計データ
D_i：y_i の観測誤差分散
$\boldsymbol{x}_i$：i 番目の小地域の共変量データ

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + v_i + \varepsilon_i, \quad i = 1, \ldots, m$$

v_i：変量効果 $\sim \mathcal{N}(0, \sigma_v^2)$　ε_i：観測誤差 $\sim \mathcal{N}(0, D_i)$
$\theta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + v_i$ とおくと，$\theta_i \sim \mathcal{N}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma_v^2)$
θ_i の Bayes 推定量：

$$\hat{\theta}_i^B(\boldsymbol{\beta}, \sigma_v^2) = E[\theta_i \mid y_i] = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \frac{\sigma_v^2}{D_i + \sigma_v^2}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})$$

$\hat{\sigma}_v^2$：σ_v^2 の推定量
$\widehat{\boldsymbol{\beta}}(\hat{\sigma}_v^2)$：$\boldsymbol{\beta}$ の GLS 推定量

$$\widehat{\boldsymbol{\beta}}(\hat{\sigma}_v^2) = \left(\sum_{i=1}^m \frac{1}{D_i + \hat{\sigma}_v^2} \boldsymbol{x}_i \boldsymbol{x}_i^\top \right)^{-1} \sum_{i=1}^m \frac{1}{D_i + \hat{\sigma}_v^2} \boldsymbol{x}_i \boldsymbol{y}_i$$

θ_i の経験最良線形不偏予測量 (EBLUP) もしくは経験ベイズ推定量：
$E[\theta_i | y_i]$ より

$$\hat{\theta}_i^{EBLUP} = \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}} + \frac{\hat{\sigma}_v^2}{D_i + \hat{\sigma}_v^2}(y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}})$$

問題点：
(a) 誤差分散 $D_1, \ldots, D_m$ は既知と仮定して解析している。実際は未知なので，推定値を与える必要がある。地域毎の集計データについては，標本平均の値は報告されているが，地域毎の標本分散の値は提供されていないので，ベイズ推定のような model-based な推測方法を使うことができない。代わりに historical data から D_i を推定してやる必要がある。
(b) 隣り合う地域も離れた地域も同じ扱いをしており，空間的な情報を用いていない。

[3] 混合効果モデルの役割

線形混合モデルと縮小推定：$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + v_i + e_i$
（データ）＝（固定効果：共通母数）＋（変量効果）＋（誤差項）
（共通母数）と（変量効果）が，安定した推定を与える役目を演ずる。

(1) 共通母数によるデータのプーリング.
$\boldsymbol{\beta}$ は全データ $(y_1, \ldots, y_m)$ の加重平均 $\widehat{\boldsymbol{\beta}}(\hat{\sigma}_v^2)$ により推定されるので，y_i の期待値は $\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}(\hat{\sigma}_v^2)$ なる安定した推定値で推定される。
母数を共通にとることによってデータをプーリングする作用が働き，結果として安定した推定が可能になる。

(2) 変量効果と縮小推定.
v_i：母数効果の場合。θ_i は y_i で推定されるので，推定誤差が問題になる。
v_i：変量効果の場合。v_i は条件付期待値 $E[v_i|y_i] = \frac{\sigma_v^2}{\sigma_v^2 + D_i}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})$ によって予測される。y_i を安定した推定値の方向へ縮小することによりリスクの改善がなされる。
こうして，線形混合モデルにおいて変量効果が，y_i を縮小する作用を生むことがわかる。

[4] 個票データを解析するためのモデル

Nested error regression model（枝分かれ誤差分散モデル）
有限母集団における各地域の平均の推定
$i = 1, \ldots, m$,　Battese, Harter, Fuller (1988) のモデル
$(Y_{i1}, \ldots, Y_{iN_i})$：第 i 郡の農作区画の総数を N_i，第 i 郡，第 j 区画の穀物の作付け面積を Y_{ij}
$\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^T$：$(Y_{i1}, \ldots, Y_{iN_i})$ から抽出された n_i 個のデータ
$\mu_i = \overline{Y}_i = N_i^{-1}(Y_{i1} + \cdots + Y_{iN_i})$ を推定したい。
Y_{ij} に線形混合モデルを当てはめる。

$$Y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + v_i + e_{ij}, \quad i = 1, \ldots, m,\ j = 1, \ldots, N_i$$

超母集団の設定：$v_i \sim \mathcal{N}(0, \sigma_v^2), e_{ij} \sim \mathcal{N}(0, \sigma_e^2)$
$\boldsymbol{Y}_i^* = (Y_{i,n_i+1}, \ldots, Y_{i,N_i})^T$：抽出されなかったデータ

$$\mu_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{N_i} + \frac{\sum_{j=n_i+1}^{N_i} Y_{ij}}{N_i} = \frac{n_i}{N_i}\overline{y}_i + \frac{\boldsymbol{j}_{N_i-n_i}^T \boldsymbol{Y}_i^*}{N_i}$$

と書けるので，$\boldsymbol{y}_i$ に基づいて $\boldsymbol{Y}_i^*$ を推定すればよい。

$\boldsymbol{y}_i$ を与えたときの $\boldsymbol{Y}_i^*$ の条件付期待値 $E[\boldsymbol{Y}_i^*|\boldsymbol{y}_i]$ を用いて $\boldsymbol{Y}_i^*$ を推定するのが自然。$(\boldsymbol{y}_i', \boldsymbol{Y}_i^{*\prime})'$ の同時密度関数は，

$$\begin{pmatrix} \boldsymbol{y}_i \\ \boldsymbol{Y}_i^* \end{pmatrix} \sim \mathcal{N}_{N_i}\left(\begin{pmatrix} \boldsymbol{x}_i \\ \boldsymbol{x}_i^* \end{pmatrix}\boldsymbol{\beta}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}\right)$$

となる。ただし，$\boldsymbol{x}_i^* = (\boldsymbol{x}_{i,n_i+1}^T, \ldots, \boldsymbol{x}_{i,N_i}^T)^T$, $\boldsymbol{\Sigma}_{11} = \sigma_e^2 \boldsymbol{I}_{n_i} + \sigma_v^2 \boldsymbol{j}_{n_i}\boldsymbol{j}_{n_i}^\top$, $\boldsymbol{\Sigma}_{12} = \sigma_v^2 \boldsymbol{j}_{n_i}\boldsymbol{j}_{N_i-n_i}^\top$, $\boldsymbol{\Sigma}_{22} = \sigma_e^2 \boldsymbol{I}_{N_i-n_i} + \sigma_v^2 \boldsymbol{j}_{N_i-n_i}\boldsymbol{j}_{N_i-n_i}^\top$ である。

$$\boldsymbol{Y}_i^*|\boldsymbol{y}_i \sim \mathcal{N}_{N_i-n_i}\left(\boldsymbol{x}_i^{*\prime}\boldsymbol{\beta} + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{y}_i - \boldsymbol{x}_i'\boldsymbol{\beta}), \boldsymbol{\Sigma}_{22} - \Sigma_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\right)$$

より，$\bar{\boldsymbol{x}}_i^* = (N_i - n_i)^{-1}\sum_{j=n_i+1}^{N_i}\boldsymbol{x}_{ij}$ とおくと，

$$E\Big[\sum_{j=n_i+1}^{N_i} Y_{ij}\Big|\boldsymbol{y}_i\Big] = E\left[\boldsymbol{j}_i^{*\prime}\boldsymbol{Y}_i^*|\boldsymbol{y}_i\right] = (N_i - n_i)\left\{\bar{\boldsymbol{x}}_i^{*\prime}\boldsymbol{\beta} + \frac{n_i\rho}{1+n_i\rho}(\bar{y}_i - \bar{\boldsymbol{x}}_i'\boldsymbol{\beta})\right\}$$

となり，観測されないデータの部分 $\sum_{j=n_i+1}^{N_i} Y_{ij}$ が

$$(N_i - n_i)\left\{\bar{\boldsymbol{x}}_i^{*\prime}\boldsymbol{\beta} + \frac{n_i\rho}{1+n_i\rho}(\bar{y}_i - \bar{\boldsymbol{x}}_i'\boldsymbol{\beta})\right\}$$

で補間されることを意味する。従って，母集団平均 $\mu_i = \bar{Y}_i$ のベイズ予測量は

$$\begin{aligned}\tilde{\mu}_i(\rho) =& E[\bar{Y}_i|\boldsymbol{y}_i] = \frac{n_i}{N_i}\bar{y}_i + \frac{1}{N_i}E\Big[\sum_{j=n_i+1}^{N_i} Y_{ij}\Big|\boldsymbol{y}_i\Big] \\ =& \bar{\boldsymbol{x}}_{i(p)}'\boldsymbol{\beta} + \left\{\frac{n_i\rho}{1+n_i\rho} + \frac{n_i}{N_i}\frac{1}{1+n_i\rho}\right\}\left(\bar{y}_i - \bar{\boldsymbol{x}}_i'\boldsymbol{\beta}\right)\end{aligned}$$

となる。ただし $\bar{\boldsymbol{x}}_{i(p)} = \sum_{j=1}^{N_i}\boldsymbol{x}_{ij}$ である。
$\rho = \sigma_v^2/\sigma_e^2$ とおくと $\boldsymbol{\beta}$ の一般化最小 2 乗推定量 (GLS) は

$$\widetilde{\boldsymbol{\beta}}(\rho) = \left(\sum_{i=1}^m \frac{n_i\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^\top}{1+n_i\rho} + \sum_{i=1}^m\sum_{j=1}^{n_i}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top\right)^{-1}\left(\sum_{i=1}^m \frac{n_i\bar{\boldsymbol{x}}_i\bar{y}_i}{1+n_i\rho} + \sum_{i=1}^m\sum_{j=1}^{n_i}\boldsymbol{x}_{ij}y_{ij}\right)$$

で与えられる。
σ_v^2, σ_e^2 はそれぞれ分散の群間成分，群内成分と呼ばれ，こうした分散成分の推定方法には Henderson の方法，Rao の MINQUE 法など様々な手法が古くから提案されてきた。
ここでは，Prasad and Rao (1990) により与えられた，Henderson (Method III) に基づいた明示的な推定方法を紹介する。

$\boldsymbol{B} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij}-\overline{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij}-\overline{\boldsymbol{x}}_i)^\top$ とおいてそのランクを r, また一般化逆行列を $\boldsymbol{B}^-$ とし，$\widetilde{\boldsymbol{\beta}} = \boldsymbol{B}^-\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij}-\overline{\boldsymbol{x}}_i)(y_{ij}-\overline{y}_i)$ とする。このとき，σ_e^2 の不偏推定量は，$N=\sum_{i=1}^{m} n_i$ に対して

$$\hat{\sigma}_e^2 = (N-m-r)^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left\{(y_{ij}-\overline{y}_i)-(\boldsymbol{x}_{ij}-\overline{\boldsymbol{x}}_i)^\top\widetilde{\boldsymbol{\beta}}\right\}^2 \tag{1}$$

となる。一方，σ_v^2 の推定については，$\boldsymbol{\beta}$ の OLS $\widehat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$ を用いて

$$\hat{\sigma}_v^{2*} = N_*^{-1}\{(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{OLS})^\top(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{OLS})-(N-p)\hat{\sigma}_e^2\}$$

で与えられる。ここで，$\boldsymbol{y}_i=(y_{i1},\ldots,y_{in_i})^\top$, $\boldsymbol{y}=(\boldsymbol{y}_1^\top,\ldots,\boldsymbol{y}_m^\top)^\top$, $\boldsymbol{X}_i=(\boldsymbol{x}_{i1}^\top,\ldots,\boldsymbol{x}_{in_i}^\top)^\top$, $\boldsymbol{X}=(\boldsymbol{X}_1^\top,\ldots,\boldsymbol{X}_m^\top)^\top$ とし，
$N_* = N-\sum_{i=1}^{m} n_i^2\overline{\boldsymbol{x}}_i^\top(\boldsymbol{X}'\boldsymbol{X})^{-1}\overline{\boldsymbol{x}}_i$ とする。$\hat{\sigma}_v^{2*}$ は負の値を取り得るので $\hat{\sigma}_v^2=\max(0,\hat{\sigma}_v^{2*})$ を用いることになる。
$\rho=\sigma_v^2/\sigma_e^2$ に $\hat{\rho}=\hat{\sigma}_v^2/\hat{\sigma}_e^2$, $\boldsymbol{\beta}$ に $\widetilde{\boldsymbol{\beta}}(\hat{\rho})$ を代入すると，母集団平均 $\mu_i=\bar{Y}_i$ の経験ベイズ予測量：

$$\tilde{\mu}_i(\hat{\rho}) = \overline{\boldsymbol{x}}'_{i(p)}\widetilde{\boldsymbol{\beta}}(\hat{\rho}) + \left\{\frac{n_i\hat{\rho}}{1+n_i\hat{\rho}}+\frac{n_i}{N_i}\frac{1}{1+n_i\hat{\rho}}\right\}\left(\overline{y}_i-\overline{\boldsymbol{x}}_i'\widetilde{\boldsymbol{\beta}}(\hat{\rho})\right)$$

縮小推定量は，n_i もしくは $\hat{\rho}$ が小さければ $\overline{y}_i$ を $\overline{\boldsymbol{x}}_i'\widetilde{\boldsymbol{\beta}}(\hat{\rho})$ の方向へ縮小することによって安定化を図っている。n_i が小さければ，データの不足を周辺もしくは全体のデータで補うことによって予測精度を高めていると解釈される。

問題点：(a) 個票データの利用の難しさ
(b) 欠損データに対応する共変量 $\boldsymbol{x}_{n_i+1},\ldots,\boldsymbol{x}_{N_i}$ のデータは利用可能か
応用例：欠損データに対して補間して予測することができる。

$$E\Big[\sum_{j=n_i+1}^{N_i} Y_{ij}\Big|\boldsymbol{y}_i\Big] = (N_i-n_i)\left\{\bar{\boldsymbol{x}}_i^{*\prime}\boldsymbol{\beta}+\frac{n_i\rho}{1+n_i\rho}(\bar{y}_i-\bar{\boldsymbol{x}}_i'\boldsymbol{\beta})\right\}$$

[5] 様々な課題と解決策
(1) 欠損データの補間

$(Y_{i1}, \boldsymbol{x}_{i1}), \ldots, (Y_{iN_i}, \boldsymbol{x}_{iN_i})$: $i = 1, \ldots, m$

$$Y_{ij} = \begin{cases} 1 & \text{if} \quad \widetilde{Y}_{ij} > 0, \\ 0 & \text{if} \quad \widetilde{Y}_{ij} \leq 0, \end{cases} \tag{2}$$

ここで $\widetilde{Y}_{ij}$ は NERM に従う。

$$\widetilde{Y}_{ij} = \boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta} + v_i + \varepsilon_{ij}, \quad i = 1, \ldots, m,\ j = 1, \ldots, N_i,$$

変量効果 v_i : $v_i \sim \mathcal{N}(0, \tau^2)$.
誤差変量 ε_{ij} : $-\varepsilon_{ij} \sim F(\cdot)$, 分布関数（正規，ロジスティック）
観測データ : $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^{\top}$, $(\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iN_i})$
欠損データ : $\boldsymbol{Y}_i^* = (Y_{in_i+1}, \ldots, Y_{iN_i})^{\top}$
予測したい量 : $\overline{Y}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} Y_{ij}$

$$\overline{Y}_i = \frac{n_i}{N_i}\overline{y}_i + \frac{Y_{i,n_i+1} + \cdots + Y_{i,N_i}}{N_i}, \quad \overline{y}_i = n_i^{-1}\sum_{j=1}^{n_i} y_{ij}$$

ベイズ予測量 :

$$\begin{aligned} \widehat{\overline{Y}}_i(\boldsymbol{\beta}, \tau^2) &= \frac{n_i}{N_i}\overline{y}_i + \frac{1}{N_i}\sum_{j=n_i+1}^{N_i} E[Y_{ij} \mid \boldsymbol{y}] \\ &= \frac{n_i}{N_i}\overline{y}_i + \frac{1}{N_i}\sum_{j=n_i+1}^{N_i} \frac{\int F(\boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta} + v_i) G(v_i \mid \boldsymbol{\beta}, \tau^2) dv_i}{\int G(v_i \mid \boldsymbol{\beta}, \tau^2) dv_i} \end{aligned}$$

ここで, v_i の確率密度関数 $g(v_i \mid \tau^2)$ に対して

$$G(v_i \mid \boldsymbol{\beta}, \tau^2) = g(v_i \mid \tau^2) \prod_{j=1}^{n_i} \left[\left\{ F(\boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta} + v_i) \right\}^{y_{ij}} \left\{ 1 - F(\boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta} + v_i) \right\}^{1-y_{ij}} \right]$$

である。$\boldsymbol{\beta}, \tau^2$ を周辺分布の対数尤度

$$\sum_{i=1}^{m} \log \left\{ \int \prod_{j=1}^{n_i} \left[\left\{ F(\boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta} + v_i) \right\}^{y_{ij}} \left\{ 1 - F(\boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta} + v_i) \right\}^{1-y_{ij}} \right] g(v_i \mid \tau^2) dv_i \right\}$$

から $\widehat{\boldsymbol{\beta}}, \hat{\tau}^2$ で推定すると，経験ベイズ予測量 $\widehat{\overline{Y}}_i^{EB} = \widehat{\overline{Y}}_i(\widehat{\boldsymbol{\beta}}, \hat{\tau}^2)$ が得られる。

(2) ベンチマーク問題と制約付きベイズ推定

$i = 1, \ldots, m,\ \boldsymbol{y} = (y_1, \ldots, y_m)^\top$

$y_i \mid \theta_i \sim \mathcal{N}(\theta_i, D_i), \quad \theta_i \sim \mathcal{N}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma_v^2)$

[1] 問題点：平均が一致しない。

m 個の地域全体の平均を経験ベイズ推定値で構成してみると，全体の標本平均の値と一致しないという問題がある

$w_i = \frac{D_i^{-1}}{\sum_{j=1}^m D_j^{-1}}$ を重みとして加重平均をとったもの $\bar{y}_w = \sum_{i=1}^m w_i y_i$ で全体の平均を推定する。

$$\sum_{i=1}^m w_i \hat{\theta}_i^{EB} \neq \sum_{i=1}^m w_i y_i$$

問題によっては，各地域の推定値の加重平均が全体の標本平均と一致することが要請されており，この要請に応えるために各地域のベイズ推定値を調整することをベンチマーク問題という。

[2] 問題点：ベイズ推定値は縮小し過ぎる傾向にある。

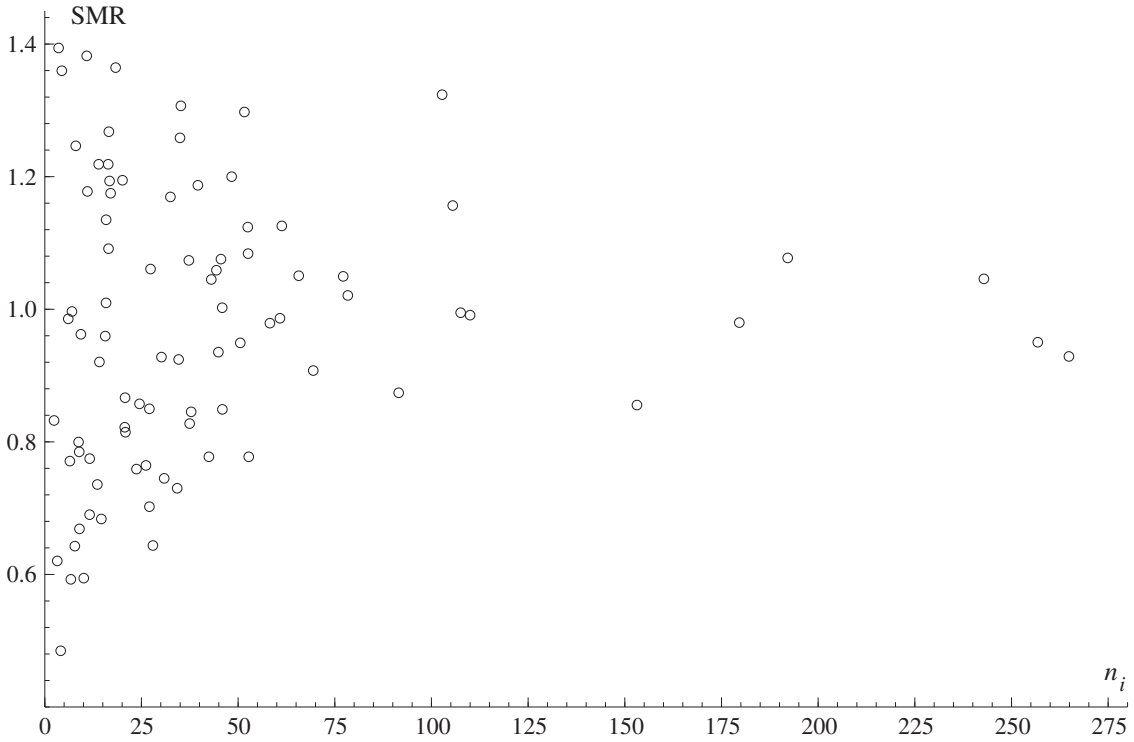

Figure: 埼玉県における胃がんによる女性死亡リスク SMR に関する市町村別プロット

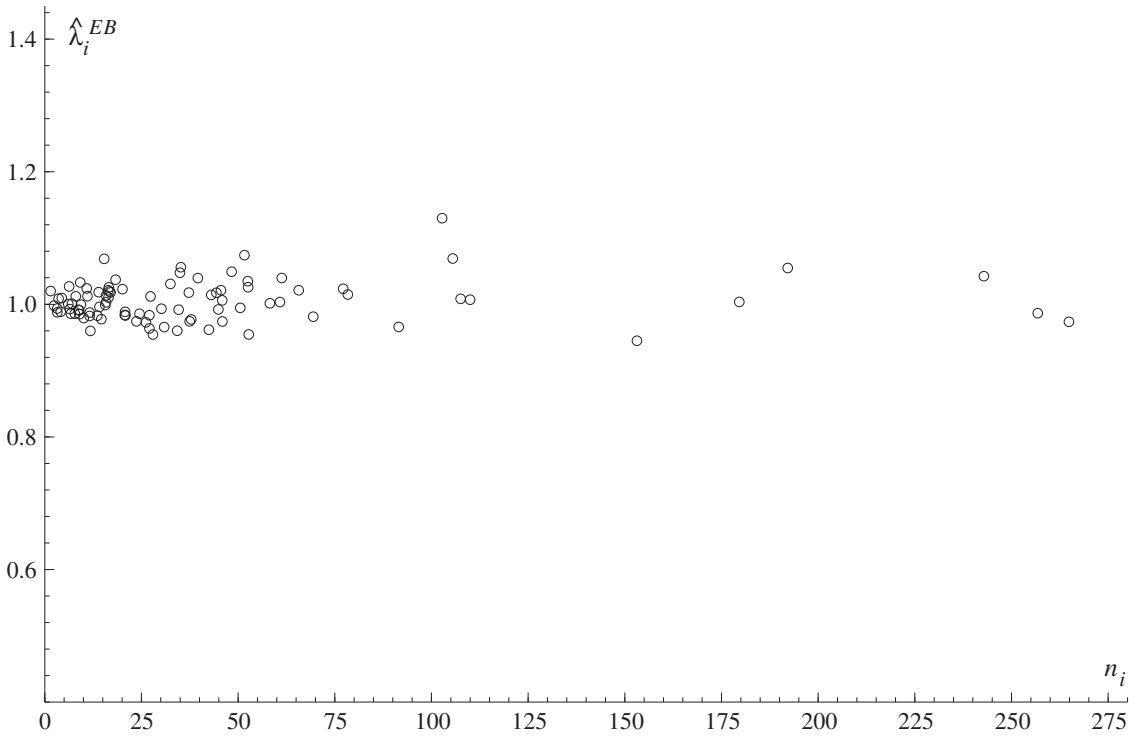

Figure: 埼玉県における胃がんによる女性死亡リスク EB に関する市町村別プロット

Louis (1984)：ベイズ推定値の平均・分散を事後分布の平均・分散と比較する。$\hat{\theta}_i^B = E[\theta_i \mid \boldsymbol{y}]$

(B1) $\sum_{i=1}^m w_i \hat{\theta}_i^B = \sum_{i=1}^m w_i E[\theta_i|\boldsymbol{y}]$

(B2) $\sum_{i=1}^m w_i \{\hat{\theta}_i^B - \bar{\hat{\theta}}_B\}^2 \leq \sum_{i=1}^m w_i E[\{\theta_i - \bar{\theta}\}^2|\boldsymbol{y}]$

ただし，$\bar{\hat{\theta}}_B = \sum_{i=1}^m w_i \hat{\theta}_i^B$，$\bar{\theta} = \sum_{i=1}^m w_i \theta_i$

実際，不等式 (B2) は次から従う。

$$\sum_{i=1}^m w_i E[\{\theta_i - \bar{\theta}\}^2|\boldsymbol{y}] = \sum_{i=1}^m w_i \left\{\hat{\theta}_i^B - \bar{\hat{\theta}}_B\right\}^2 + \sum_{i=1}^m w_i Var(\theta_i - \bar{\theta}|\boldsymbol{y})$$

$$\geq \sum_{i=1}^m w_i \left\{\hat{\theta}_i^B - \bar{\hat{\theta}}_B\right\}^2$$

この不等式は，ベイズ推定値の分散が事後分布の分散より小さいことを意味している。

[3] 解決策：制約付きベイズ推定法

(平均制約) $\sum_{i=1}^m w_i \hat{\theta}_i^{CB} = \sum_{i=1}^m w_i y_i$

(分散制約) $\sum_{i=1}^m w_i \{\hat{\theta}_i^{CB} - \bar{\hat{\theta}}_{CB}\}^2 = \sum_{i=1}^m w_i \{\widehat{\mu}_i^B - \bar{\hat{\theta}}_B\}^2 + \sum_{i=1}^m w_i (1 - w_i) \frac{D_i \sigma_v^2}{D_i + \sigma_v^2}$

ラグランジュの未定乗数法で解く。

$$L(\hat{\theta}_1^{CB},\ldots,\hat{\theta}_m^{CB},\lambda_1,\lambda_2) = \sum_{i=1}^{m} E[(\hat{\theta}_i^{CB}-\theta_i)^2|\boldsymbol{y}] - \lambda_1\Big\{\sum_{i=1}^{m} w_i\hat{\theta}_i^{CB} - \sum_{i=1}^{m} w_i y_i\Big\}$$
$$- \lambda_2\Big\{\sum_{i=1}^{m} w_i\{\hat{\theta}_i^{CB} - \bar{\hat{\theta}}_{CB}\}^2 - \sum_{i=1}^{m} w_i\{\widehat{\mu}_i^{B} - \bar{\hat{\theta}}_{B}\}^2 - \sum_{i=1}^{m} w_i(1-w_i)\frac{D_i\sigma_v^2}{D_i+\sigma_v^2}\Big\}$$

の最小化問題を解いたものを制約付きベイズ推定量という。

$$\hat{\theta}_i^{CB} = \hat{\theta}_i^{B} + \{a_B - 1\}\{\hat{\theta}_i^{B} - \sum_{j=1}^{m} w_j\hat{\theta}_j^{B}\} + \sum_{i=1}^{m} w_i(y_i - \hat{\theta}_i^{B})$$

ここで，$\Delta_v = \sum_{i=1}^{m} w_i(1-w_i)\frac{D_i\sigma_v^2}{D_i+\sigma_v^2}$ に対して

$$\{a_B\}^2 = 1 + \frac{\Delta_v}{\sum_{i=1}^{m} w_i\{\hat{\theta}_i^{B} - \sum_{j=1}^{m} w_j\hat{\theta}_j^{B}\}^2}$$

更なる問題点：$y_i > 0$ のときには変換が必要。制約条件 $\hat{\theta}_i^{CB} > 0$ を課す必要。Ghosh, Kubokawa and Kawakubo(2015)

利点：モデルの間違った設定に対する頑健性を与えている。

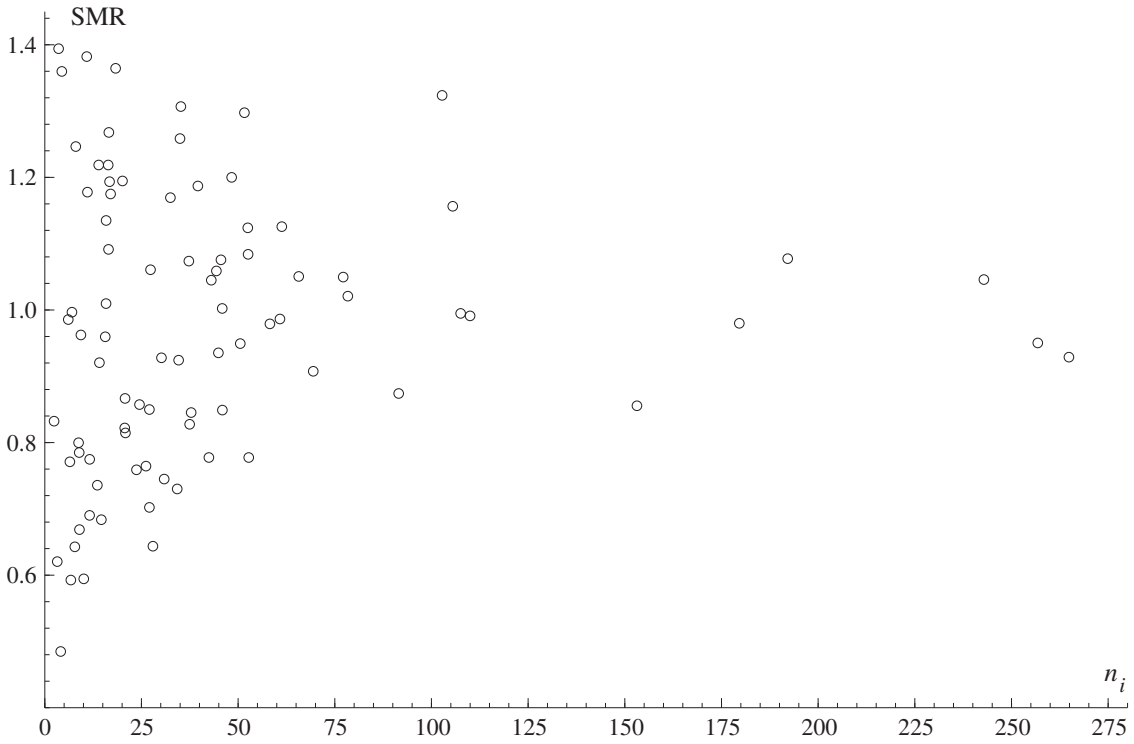

Figure: 埼玉県における胃がんによる女性死亡リスク SMR に関する市町村別プロット

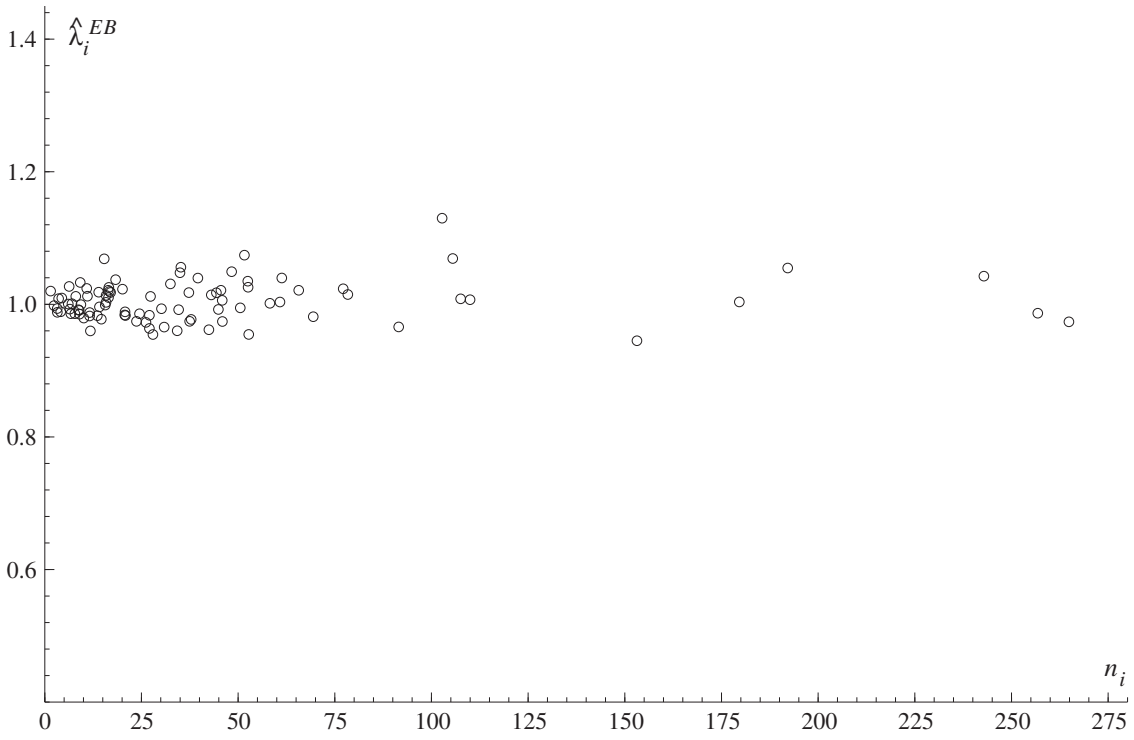

Figure: 埼玉県における胃がんによる女性死亡リスク EB に関する市町村別プロット

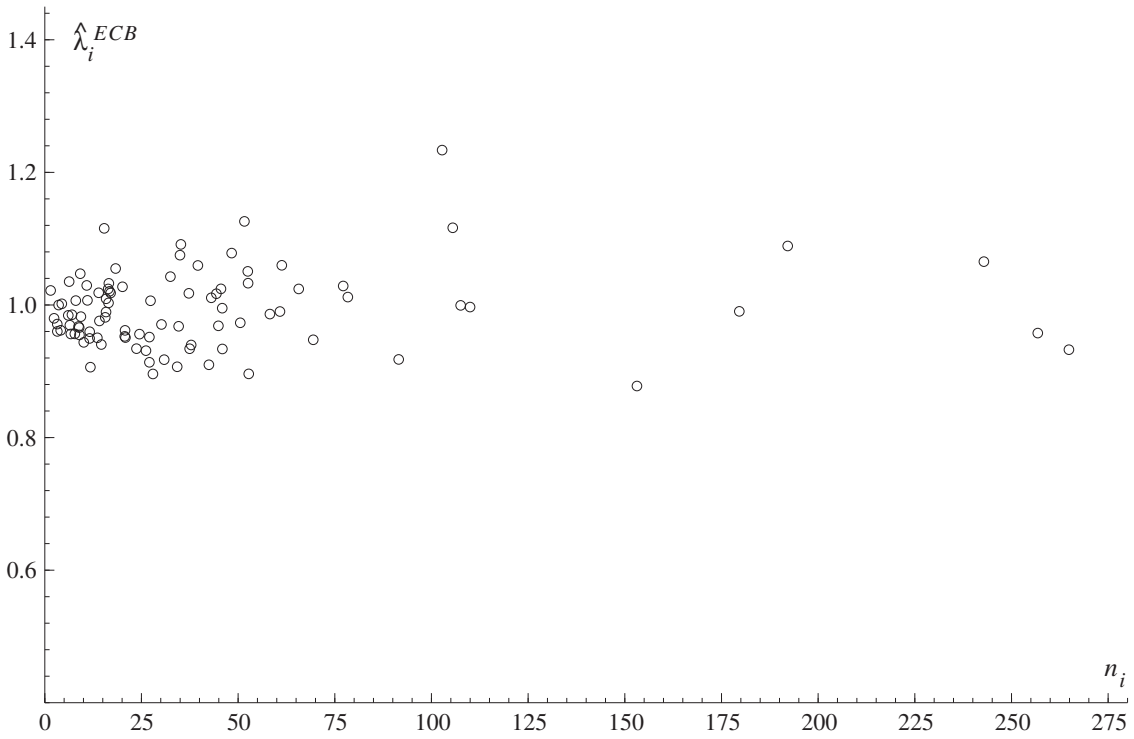

Figure: 埼玉県における胃がんによる女性死亡リスク CEB に関する市町村別プロット

(3) 不均一分散をもつモデルへの拡張

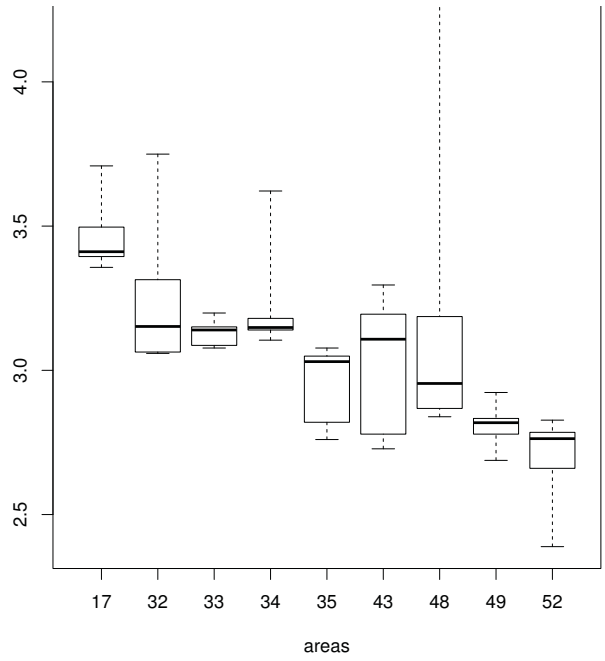

Figure: Boxplots of the Posted Land Price Data for Selected Areas

地域レベルモデル：$i = 1, \ldots, m$, $\boldsymbol{y} = (y_1, \ldots, y_m)^\top$
$y_i \mid \theta_i \sim \mathcal{N}(\theta_i, \sigma_i^2)$, $\quad \theta_i \sim \mathcal{N}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma_v^2)$

S_i^2 を σ_i^2 の推定量とする。$n_i S_i^2 / \sigma_i^2 \sim \chi_{n_i}^2$

問題点：n_i が小さいとき，S_i^2 は一致性がない。予測量に一致性がない。
解決策：$\sigma_1^2, \ldots, \sigma_m^2$ に制約を入れる。
(a) 等号制約（強い制約）：$\sigma_1^2 = \cdots = \sigma_m^2 = \sigma^2$
(b) 変量分散モデル：$\sigma_i^{-1} \sim Gamma(a, b)$, (a, b)：未知
近似ベイズ推定量：平均と分散の 2 重縮小推定 Tamae and Kubokawa (2015)

$$\hat{\theta}_i^{AB} = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + (n_i S_i^2 + 2/b)/(n_i + a/2)} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})$$

(c) 分散関数モデル：Sugasawa and Kubokawa (2017)

$$\sigma_i^2 = g(\boldsymbol{z}_i^\top \boldsymbol{\gamma})$$

例：$g(\boldsymbol{z}_i^\top \boldsymbol{\gamma}) = (\boldsymbol{z}_i^\top \boldsymbol{\gamma})^2$, $g(\boldsymbol{z}_i^\top \boldsymbol{\gamma}) = \exp\{\boldsymbol{z}_i^\top \boldsymbol{\gamma}\}$

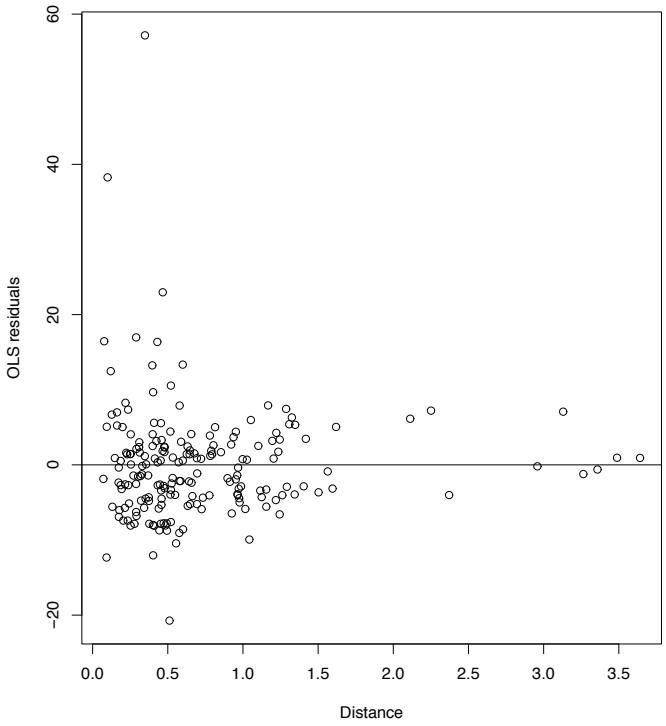

Figure: Plot of OLS Residuals Against Distance D_{ij}

(4) 変換を伴う線形混合モデル

支出，収入や生産量など観測データは正の値をとることが多い。この場合，対数変換したものに線形混合モデルを当てはめることが一般になされているが，対数変換が必ずしも相応しいとは限らない。

(I) 変換関数：$y > 0$ なので，$h(\cdot) : (0, \infty) \to (-\infty, \infty)$

(1) Box-Cox 変換

$$h_\lambda^{BC}(y) = \begin{cases} (y^\lambda - 1)/\lambda, & \lambda \neq 0, \\ \log y, & \lambda = 0, \end{cases}$$

変換パラメータ λ の MLE は一致性をもたない。

(2) 双巾変換 (Dual Power Transformation)

$$h_\lambda^{DP}(y) = \begin{cases} (y^\lambda - y^{-\lambda})/2\lambda, & \lambda > 0, \\ \log y, & \lambda = 0. \end{cases}$$

変換パラメータ λ の MLE は一致性をもつ。

(II) モデルの作り方：$y_1, \ldots, y_m$, $y_i > 0$

(1) Fay-Herriot モデル：$i = 1, \ldots, m$

$$
\begin{aligned}
y_i &= \theta_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, D_i), \\
\theta_i &= \boldsymbol{x}_i^\top \boldsymbol{\beta} + v_i, \quad v_i \sim \mathcal{N}(0, \sigma_v^2)
\end{aligned}
$$

$\theta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + v_i$ の予測量：$\hat{\theta}_i^B = \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}} + \frac{\hat{\sigma}_v^2}{D_i + \hat{\sigma}_v^2}(y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}})$
$y_i > 0$ なので相応しくない。

(2) 変換モデル：$y_i > 0$ のとき，y_i を変換 $h(y_i, \lambda)$

$$
h_\lambda(y_i) = \theta_i + \varepsilon_i, \ i = 1, \ldots, m
$$

(a) θ_i の予測量を逆変換したもの

$$
\hat{\theta}_i^B(h_\lambda) = \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}} + \frac{\hat{\sigma}_v^2}{D_i + \hat{\sigma}_v^2}(h_\lambda(y_i) - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}})
$$

に対して $h_\lambda^{-1}(\hat{\theta}_i^B(h_\lambda))$ を考える。Sugasawa and Kubokawa (2015)
逆変換は何を予測していることになるのか。

(b) θ_i を逆変換したものの予測量：Sugasawa and Kubokawa (2017)

$$
E[h_\lambda^{-1}(\theta_i)|y_i] = E[h_\lambda^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta} + v_i)|y_i] = \int_{-\infty}^{\infty} h_\lambda^{-1}(t)\phi(t; \hat{\theta}_i^B(h_\lambda), \sigma_i^2)dt
$$

ただし $\sigma_i^2 = \sigma_v^2 D_i/(\sigma_v^2 + D_i)$, $\phi(t; \hat{\theta}_i, \sigma_i^2) \sim \mathcal{N}(\hat{\theta}_i, \sigma_i^2)$

(3) 個票データ解析のための変換モデル：有限母集団の枠組み
$Y_{ij} > 0, \quad i = 1, \ldots, m, \ j = 1, \ldots, N_i$
母集団平均 $\overline{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} Y_{ij}$ を推定したい。
変換：$h_\lambda(\cdot) : (0, \infty) \to (-\infty, \infty) \quad h_\lambda(Y_{ij}) = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + v_i + e_{ij}$
超母集団の設定：$v_i \sim \mathcal{N}(0, \sigma_v^2), e_{ij} \sim \mathcal{N}(0, \sigma_e^2)$
$\boldsymbol{y}_i = (y_{i,1}, \ldots, y_{i,n_i})^\top$：抽出されたデータ
$\boldsymbol{Y}_i^* = (Y_{i,n_i+1}, \ldots, Y_{i,N_i})^\top$：抽出されなかったデータ
$\overline{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} Y_{ij}$ の予測量：

$$
E[\overline{Y}_i \mid \boldsymbol{y}_i] = \frac{\sum_{j=1}^{n_i} y_{ij}}{N_i} + \frac{\sum_{j=n_i+1}^{N_i} E[Y_{ij} \mid \boldsymbol{y}_i]}{N_i} = \frac{n_i}{N_i}\overline{y}_i + \frac{\sum_{j=n_i+1}^{N_i} E[Y_{ij} \mid \boldsymbol{y}_i]}{N_i}
$$

ここで $E[Y_{ij}|\boldsymbol{y}_i] = E[h_\lambda^{-1}(\boldsymbol{x}_{ij}^T \boldsymbol{\beta} + v_i + e_{ij})|\boldsymbol{y}_i]$ を計算する必要がある。

国別地域別の貧困指数 (poverty indices) の推定への応用

(4)Unmatched Sampling Linking Model： Sugasawa, Kubokawa and Rao (2017)

$$y_i = \theta_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, D_i),$$
$$h(\theta_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} + v_i, \quad v_i \sim \mathcal{N}(0, \sigma_v^2)$$

$\theta_i > 0$, $h(\cdot) : (0, \infty) \to (-\infty, \infty)$
例えば， $h(\theta_i) = \log \theta_i$, $h(\theta_i) = \log\{\theta_i/(1-\theta_i)\}$
$\theta_i = h^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta} + v_i)$ の予測量

$$E[\theta_i | y_i] = E[h^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta} + v_i) | y_i] = \int \theta_i \pi(\theta_i | y_i) d\theta_i,$$

$$\pi(\theta_i | y_i) = \frac{h'(\theta_i) \exp\{-(y_i - \theta_i)^2/2D_i - (h(\theta_i) - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2/2\sigma_v^2\}}{\int h'(\theta_i) \exp\{-(y_i - \theta_i)^2/2D_i - (h(\theta_i) - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2/2\sigma_v^2\} d\theta_i}$$

(5)　離散データ解析

地域別の疾病数，死亡数，犯罪数などの計数データの解析について，地域別変量効果を導入したモデルを用いることによって精度の高い地域別推定値を与えることができる。ポアソン・ガンマモデル，2 項・ベータモデルなど，指数型分布族の中で分散が平均の 2 次関数で表現できている分布族については共役な事前分布を用いて経験ベイズ推定値を求めることができ，安定した地域別推定値を与えることができる。また指数型分布族の平均にリンク関数を設定しそこに事前分布として正規分布を想定するモデルは一般化線形混合モデル (GLMM) と呼ばれいる。

(6)　パネルデータ解析

地域別なデータが時系列的に取られているとき，いわゆるパネルデータの分析に地域効果を組み入れて解析するモデルも提唱されている。母数の識別可能性や推定方法が問題になる。

(7)　変数選択規準

線形混合モデルの変数選択については，通常の AIC, BIC とは異なり，変量効果に基づいた予測誤差の改善を考慮する規準として条件付き AIC が提案され，Kawakubo and Kubokawa (2014) などで更なる展開がなされている。

(8) 予測誤差の評価

EBLUP のような縮小推定量は推定精度を高めるために導入されているので，実際予測誤差がどの程度改善されているのかを見積もる必要がある。そこで EBLUP の平均 2 乗誤差 (MSE) の $m \to \infty$ のもとでの漸近展開と，その 2 次漸近不偏推定量の導出を行う。

予測誤差を評価する別の方法は予測信頼区間を作ることで，$m \to \infty$ のもとで 2 次漸近的に信頼係数 $1 - \gamma$ に一致する信頼区間の構成を行う。

このような 2 次不偏などの手法を導出するには，テーラー展開に基づいた解析的な方法とブートストラップに基づいた方法が一般的である。

解析的な方法によると，2 次不偏推定量を明示的に求めることができるが，簡単なモデルについては計算ができても複雑なモデルになると偏微分等の計算が困難になる。これを回避する一つのアプローチは数値微分を用いるもので，2 次不偏性を理論上保証しつつ偏微分の複雑な計算をする必要なく計算機の力を借りて求めることができる。

ブートストラップ法もくしはパラメトリック・ブートストラップ法は容易に計算できるので便利であるが，ML などの繰り返し計算を必要とする推定手法が含まれていると，相当な計算時間がかかってしまうという欠点がある。

匿名性を確保した所得等内訳情報の作成

星野　伸明
金沢大学・経済学経営学系

本研究は科研費の助成を受けている。

動機付け：統計委員会諮問第 76 号の答申より

- 「国民生活基礎調査に係る匿名データの作成について（平成２７年１月２９日）」から引用：
  - 前回答申（諮問第 54 号、平成 25 年 9 月 27 日答申）における「今後の課題」への対応
    - ＊【所得票の内訳情報の提供】本計画では、所得等の情報について、総額についてのみトップコーディングして提供することとしている。これに対して、所得等の内訳情報も有用性が高いことからトップコーディング等の匿名化措置をして提供することが検討されたが、総額と内訳の情報に整合性が取れないこと、匿名性が十分に確保することができないことが明らかになった。　このような検討の結果、所得等は総額のみをトップコーディングして提供することとし、その内訳情報の提供については、今後、より精緻な匿名化手法に関する慎重な研究・検討が必要であることから、引き続き、後述「３　今後の課題」で示した方向で検討する必要がある。
  - 今後の課題
    - ＊【所得票の内訳情報の提供】本計画においては、所得票に含まれる情報について

世帯の総所得、課税等の状況及び掛金に限定して提供することとしている。しかしながら、近年、社会保障や所得格差等に関する研究の重要性が増しており、その分析には所得等に関する内訳情報の必要性が指摘されている。一方、匿名性を十分に確保した内訳情報のデータ作成方法は、確立されておらず、より精緻な匿名化手法に関する慎重な研究・検討が必要となっている。　このため、今後、所得等の内訳情報の提供に向け、匿名性と有用性の確保の観点から、トップコーディング以外の適用も含めて匿名化措置を検討する必要がある。

- （潜在的には国民生活基礎調査に限らず）匿名データの作成において、匿名性を確保した所得等内訳情報の作成が必要とされている。
  - しかし方法について研究不足である。

問題の匿名化措置

- 国民生活基礎調査の所得票（匿名データ）
  - 総額のみをトップコーディングして提供；
    1. 総所得：2200 万円以上、1 万円単位
    2. 拠出金合計（税金＋社会保険料）：490 万円以上、1 千円単位
    3. 掛金（企業年金・個人年金等掛金）：80 万円以上、1 千円単位
- 同調査票 (H29)
  - 所得の内訳：雇用者所得、事業所得、農耕・畜産所得、家内労働所得、財産所得、公的年金・恩給、雇用保険、児童手当等、その他の社会保障給付金、仕送り、企業年金・個人年金等、その他の所得
  - 所得は全て 1 万円単位、5 桁か 4 桁。
  - 拠出金や掛け金は 1 千円単位、5 桁。

そもそも論

1. 所得は匿名化措置を施すべきか。
   - 高額所得者は目立つかもしれないので（識別に使える）キー変数かもしれない。
   - キー変数でなくてもセンシティブ変数であろう。
   - キー変数なら措置は妥当。センシティブ変数に措置をする必要はないが、措置は安心につながる。
2. 何故トップコーディング？
   - 裾の個体は目立ちやすいので、キーとして使えなくしたい。
   - 個体量が少ない領域をまとめると識別可能な個体量は減りやすい。
3. トップコーディング時の内訳整合性をどのように考えるか。
   - 内訳は比（0.1%単位でどうか）なら整合的。桁丸めによる保護効果有り。
     - トップコーディングされない個体も内訳を比で与えると場合分け要らず。

内訳を比で与えることの論点

1. 内訳の比ではなく実額が知りたい？
   - 総額がトップコーディングされている場合、内訳の実額が分からない。
   - トップコーディングでなければウソの総額を与える（攪乱）しかないと思われる。
2. 構成比が特異な個体は安全？
   - 構成比はキー変数ではないと思うが、丸めの桁数で保護効果が変わる。
   - 構成比を攪乱すれば保護効果は上がる。
3. 内訳の総和が 100%にならないことを許す？
   - 構成比の下 1 桁が四捨五入だと総和は 1 になるとは限らない。
   - 総和 1 が所与で期待値が真の構成比になるような確率ベクトルを用いる “random rounding” がスマート。
   - random rounding は攪乱の一種。

(Naive) Random Rounding

- 各構成比の「有効桁数」を 10^{-d} とおく ($d = 0, 1, 2, \ldots$)。
- 真の第 j 構成比が π_j の時、π_j を 10^{-d} で割った余りを r_j と書く ($j = 1, 2, \ldots, J$)。
- 余りの総和を $r_\cdot := \sum_{j=1}^{J} r_j$ と書く。$r_\cdot$ は 10^{-d} か 0 のどちらか。以下 0 でないとする。
  - 理論的には 10 進法にこだわる必要がない。$r_\cdot$ が 0 にならない底を選べる。
  - $\pi_j > 0$ となる項目全てについて $r_j > 0$ となる底を選ぶべき。
    - $\pi_j = 0$ となる項目を保護するには歪みを入れるしかない。
- 確率変数ベクトル $\boldsymbol{X}_J := (X_1, X_2, \ldots, X_J)$ は、第 j セル確率が $r_j / r_\cdot =: p_j$ で 10 を配る多項分布 $\mathtt{Multi}(10, \boldsymbol{p})$ に従うとする。
- π_j を 10^{-d} で割った商に 10^{-d} をかけて $X_j \times 10^{-d-1}$ を加えた値を、丸められた第 j 構成比 Y_j とする。
- 明らかに $\mathsf{E}(Y_j) = \pi_j$ かつ $\sum_{j=1}^{J} Y_j = 1$ となる。

一般化：構成比の攪乱

- Random Rounding だと下 1 桁のみランダムだが、1 桁である必要はない。
- 前のページで $\boldsymbol{X}_J \sim \mathtt{Multi}(10, \boldsymbol{p})$ のところを $\mathtt{Multi}(10^c, \boldsymbol{p})$ として Y_j を調整すれば下 c 桁をランダムに出来る ($c = 0, 1, 2, \ldots$)。
- 一般に任意の総桁・有効桁で

$$\mathsf{E}(Y_j) = \pi_j, \quad \sum_{j=1}^{J} Y_j = 1$$

  を満たす非負の構成比をランダムに生成可能。
  - 攪乱値 Y_j は有限桁なので離散変数。
- この手法の安全性は多項分布のランダムネスに依存する。

攪乱の安全性

- 攪乱の安全性についての考え方は諸説あるが、ここでは公表値 $\boldsymbol{x}$ から真値 $\boldsymbol{p}$ を推測できる精度と考える。
- 有名な定式化では、正の ϵ について $\forall(\boldsymbol{x}, \boldsymbol{p})$

$$\frac{\mathsf{P}(\boldsymbol{X} = \boldsymbol{x}; \boldsymbol{p} + \Delta)}{\mathsf{P}(\boldsymbol{X} = \boldsymbol{x}; \boldsymbol{p})} \leq \exp(\epsilon) \tag{1}$$

  を「ϵ– 差分プライベート (DP)」と呼ぶ。ただし Δ は真値 $\boldsymbol{p}$ の 1 単位変化を表す。
- (1) 式の対数をとると $|\Delta| = 1$ なので

$$\frac{\log \mathsf{P}(\boldsymbol{X} = \boldsymbol{x}; \boldsymbol{p} + \Delta) - \log \mathsf{P}(\boldsymbol{X} = \boldsymbol{x}; \boldsymbol{p})}{|\Delta|} \leq \epsilon \tag{2}$$

- 差分プライバシーはもともとデータベースの一個体の変化を考えているので $\boldsymbol{p}$ が離散的に変化する。
- 我々は (2) 式で $|\Delta| \to 0$ として、対数尤度関数の傾きが ϵ 以下なら安全と考えよう。

母数の秘匿

- $\mathsf{P}(\boldsymbol{X} = \boldsymbol{x}; \boldsymbol{p})$ が 0 になる $(\boldsymbol{x}, \boldsymbol{p})$ が存在したら差分プライベートではない。
  - 多項分布は空セル（$p_j = 0$）で $\mathsf{P}(X_j > 0) = 0$ なので差分プライベートではない。
- 全てのセル確率が正になるように多項分布の母数を秘匿：$\tilde{p_j} = p_j + \delta_j$
- Δ が p_j の微小変化とみなす：

$$\begin{aligned}
&(\log \mathsf{P}(\boldsymbol{X} = \boldsymbol{x}; \tilde{\boldsymbol{p}} + \Delta) - \log \mathsf{P}(\boldsymbol{X} = \boldsymbol{x}; \tilde{\boldsymbol{p}}))/|\Delta| \\
&= \quad (x_j(\log(p_j + \delta_j + \Delta) - \log(p_j + \delta_j)))/|\Delta| \to x_j/\tilde{p}_j
\end{aligned}$$

- 従って差分プライベートであるためには、$\forall(x_j, p_j)$

$$x_j/\tilde{p}_j \leq \epsilon$$

- つまり $\tilde{p}_j$ が最小となる j について $x_j = 10^c$ であっても上式の成立を要求。
  - 後で母数の総和が 1 という条件を入れて修正する。
- 結局、$\tilde{p}_j$ が一様かつ c が小さいほど安全。

多項分布の一般化

- 構成比の攪乱では $\boldsymbol{X}_J \sim \texttt{Multi}(10^c, \boldsymbol{p})$ とした結果、$\mathsf{E}(Y_j) = \pi_j, \sum_{j=1}^{J} Y_j = 1$ を満たす非負の構成比がランダムに生成された。
- この攪乱の安全性は多項分布の確率関数に依存して評価された。分布が変われば安全性も変わる。
- 多項分布と値域が同じで周辺の期待値も同じ分布に $\boldsymbol{X}_k$ が従うなら、同じ制約を満たす Y_j が生成可能。
- 以下では所与の π について、$\boldsymbol{F}_J \in \mathcal{F}_{|n,J} := \{\boldsymbol{f}_J : f_j \in \mathbb{N}_0, j \in [J], \sum f_j = n\}$ かつ $\mathsf{E}(F_j) = n\pi_j$ となる分布族を構成してその性質を示す。

準備）ベル多項式

- (Total) Bell polynomial:

$$B_n(\boldsymbol{w}) := n! \sum_{\boldsymbol{s} \in \mathcal{S}_{|n}} \prod_{i=1}^{n} \left(\frac{w_i}{i!}\right)^{s_i} \frac{1}{s_i!},$$

where

$$\mathcal{S}_{|n} := \{\boldsymbol{s} : s_i \in \mathbb{N}_0, i \in \mathbb{N}, \sum_{i=1}^{\infty} i s_i = n\}.$$

- $B_0(\cdot) := 1.$
- Partial Bell polynomial:

$$B_{n,k}(\boldsymbol{w}) := n! \sum_{\boldsymbol{s} \in \mathcal{S}_{|n,k}} \prod_{i=1}^{n} \left(\frac{w_i}{i!}\right)^{s_i} \frac{1}{s_i!},$$

where

$$\mathcal{S}_{|n,k} := \{\boldsymbol{s} : s_i \in \mathbb{N}_0, i \in \mathbb{N}, \sum_{i=1}^{\infty} i s_i = n, \sum_{i=1}^{\infty} s_i = k\}.$$

- Simple fact: $B_n(\boldsymbol{w}) = \sum_{k=0}^{n} B_{n,k}(\boldsymbol{w})$.
- We write $\lambda\boldsymbol{w} = (\lambda w_1, \lambda w_2, \lambda w_3, \ldots)$. $\Rightarrow B_n(\lambda\boldsymbol{w}) = \sum_{k=0}^{n} \lambda^k B_{n,k}(\boldsymbol{w})$.
- $\Rightarrow dB_n(\lambda\boldsymbol{w})/d\lambda = \sum_{k=0}^{n} k\lambda^{k-1} B_{n,k}(\boldsymbol{w})$
- If $w_i \geq 0, i \in [n]$, then $B_n(\boldsymbol{w}) \geq 0 \Rightarrow dB_n(\lambda\boldsymbol{w})/d\lambda \geq 0$ if $\lambda \geq 0$.

Bell Polynomial Distribution

- Define, for $n \in \mathbb{N}, J \in \mathbb{N}, \lambda_j \geq 0, j \in [J], w_j \geq 0, j \in [n]$, the J dimensional Bell polynomial distribution with parameters $(n, \lambda_1, \lambda_2, \ldots, \lambda_J, \boldsymbol{w})$ by

$$p(\boldsymbol{f}_J) = \binom{n}{\boldsymbol{f}_J} \frac{1}{B_n(\lambda_\cdot \boldsymbol{w})} \prod_{j=1}^{J} B_{f_j}(\lambda_j \boldsymbol{w}), \quad \boldsymbol{f}_J \in \mathcal{F}_{|n,J}. \tag{3}$$

- Denote this distribution by $\texttt{BellP}_J(n, \lambda_1, \lambda_2, \ldots, \lambda_J, \boldsymbol{w})$.
- When $\boldsymbol{w} = (1, 0, 0, 0, \ldots)$, it reduces to the J dimensional multinomial distribution with cell probabilities $\lambda_j/\lambda_\cdot, j \in [J]$.
- When $\sum_{i=1}^{\infty} w_i/i! < \infty$, it reduces to the Conditional Compound Poisson (CCP) distributions (H, 2009).

Marginal Moments

- Write $\pi_j := \lambda_j/\lambda_.$.
- **Theorem 1** *Suppose that* $\boldsymbol{F}_J \sim \texttt{BellP}_J(n, \lambda_1, \lambda_2, \ldots, \lambda_J, \boldsymbol{w})$. *Then*

$$\mathsf{E}(F_j) = n\pi_j, \quad j \in [J]. \tag{4}$$

$$\mathsf{V}(F_j) = n\pi_j(1-\pi_j)\phi(n, \lambda_., \boldsymbol{w}), \quad j \in [J], \tag{5}$$

where

$$\phi(n, \lambda_., \boldsymbol{w}) = 1 + \frac{\lambda_.(n-1)!}{B_n(\lambda_.\boldsymbol{w})} \sum_{i=0}^{n-2} \frac{B_i(\lambda_.\boldsymbol{w}) w_{n-i}}{i!(n-i-2)!}.$$

$$\mathsf{Cov}(F_i, F_j | N = n) = -n\pi_i\pi_j\phi(n, \lambda_., \boldsymbol{w}), \quad i \in [J], j \in [J], i \neq j. \tag{6}$$

- **Remark 1** $\phi(n, \lambda_., \boldsymbol{w}) = 1$ *if and only if* $w_i = 0$ *for* $i \geq 2$.
- **Corollary 1** *For all* $\boldsymbol{w}$ *the correlation matrix of* $\texttt{BellP}_J(n, \lambda_1, \lambda_2, \ldots, \lambda_J, \boldsymbol{w})$ *is that of* $\texttt{Multi}(n, \pi_1, \pi_2, \ldots, \pi_J)$.

ex) Negative Hypergeometric Distribution

- $w_i = (i-1)!, \quad B_n(\lambda\boldsymbol{w}) = \lambda(\lambda+1)\cdots(\lambda+n-1)$.
- $\texttt{BellP}_J(n, \alpha_1, \alpha_2, \ldots, \alpha_J, (0!, 1!, 2!, \ldots))$'s pmf:

$$p(\boldsymbol{f}_J) = \binom{n}{\boldsymbol{f}_J} \frac{\Gamma(\alpha_.)}{\Gamma(\alpha_. + n)} \prod_{j=1}^{J} \frac{\Gamma(\alpha_j + f_j)}{\Gamma(\alpha_j)}, \quad \boldsymbol{f}_J \in \mathcal{F}_{|n,J}.$$

- Write $\pi_j = \alpha_j/\alpha_.$.
- When $\boldsymbol{F}_J \sim \texttt{BellP}_J(n, \alpha_1, \alpha_2, \ldots, \alpha_J, (0!, 1!, 2!, \ldots))$, it is known that

$$\mathsf{E}(F_j) = n\pi_j.$$

$$\mathsf{V}(F_j) = n\pi_j(1-\pi_j)(1 + (n-1)/(\alpha_. + 1)).$$

- Hence $\phi(n, \alpha_., (0!, 1!, 2!, \ldots))$ must be $1 + (n-1)/(\alpha_. + 1)$.
- **Proposition 1**

$$\frac{(n-1)!\alpha_.}{\Gamma(\alpha_. + n)} \sum_{i=0}^{n-2} \frac{(n-i-1)\Gamma(\alpha_. + i)}{\Gamma(i+1)} = \frac{n-1}{\alpha_. + 1}.$$

Direct Proof of Proposition 1

- **Lemma 1**

$$\sum_{i=0}^{n} \frac{\Gamma(a+i)}{\Gamma(1+i)} = \frac{\Gamma(1+n+a)}{a\Gamma(1+n)}.$$

- Thus we have

$$\begin{aligned}
\phi(n, \alpha., (0!, 1!, 2!, \ldots)) &= 1 + \frac{(n-1)!\alpha.}{\Gamma(\alpha.+n)} \sum_{i=0}^{n-2} \frac{(n-i-1)\Gamma(\alpha.+i)}{\Gamma(i+1)} \\
&= 1 + \frac{(n-1)!\alpha.}{\Gamma(\alpha.+n)} \left\{ (n-1) \sum_{i=0}^{n-2} \frac{\Gamma(\alpha.+i)}{\Gamma(i+1)} - \sum_{i=1}^{n-2} \frac{\Gamma(\alpha.+i)}{\Gamma(i)} \right\} \\
&= 1 + \frac{(n-1)!\alpha.}{\Gamma(\alpha.+n)} \left\{ (n-1) \frac{\Gamma(n-1+\alpha.)}{\alpha.\Gamma(n-1)} - \frac{\Gamma(n-1+\alpha.)}{(\alpha.+1)\Gamma(n-2)} \right\} \\
&= 1 + \frac{(n-1)!\alpha.}{\Gamma(\alpha.+n)} \left\{ \frac{\Gamma(n+\alpha.)}{\alpha.(\alpha.+1)\Gamma(n-1)} \right\} \\
&= 1 + (n-1)/(\alpha.+1). \qquad \square
\end{aligned}$$

ex) Quasi-Multinomial

- $\texttt{BellP}_J(n, \theta_1/\lambda, \theta_2/\lambda, \ldots, \theta_J/\lambda, (1^0, 2^1, 3^2, \ldots))$ is "Quasi-Multinomial (type 2)" (Consul and Mittal, 1977):

$$\mathsf{P}(\boldsymbol{F}_J = \boldsymbol{f}_J) = \binom{n}{\boldsymbol{f}_J} \frac{1}{\theta.(\theta. + n\lambda)^{n-1}} \prod_{j=1}^{J} \theta_j(\theta_j + f_j\lambda)^{f_j - 1}, \quad \boldsymbol{f}_J \in \mathcal{F}_{|n,J}. \tag{7}$$

- Reparameterize (7) as $\beta := \lambda/\theta. \geq 0, \pi_j = \theta_j/\theta., j \in [J]$.

- Consul and Mittal (1977) derive

$$\mathsf{V}(F_j) = n\pi_j \left[\left\{ \frac{(n-1)!}{1+n\beta} \sum_{i=2}^{n} \frac{\pi_j + i\beta}{(n-i)!} \left(\frac{\beta}{1+n\beta} \right)^{i-2} \right\} + 1 - n\pi_j \right]. \tag{8}$$

- Theorem 1 tells us that π's can be separated in (8).

Proposition 2 *For $\beta > 0$ eq. (8) can be rewritten as*

$$V(F_j) = n\pi_j(1-\pi_j)\phi(n, \beta^{-1}, (1^0, 2^1, 3^2, \ldots)), \tag{9}$$

where

$$\phi(n, \beta^{-1}, (1^0, 2^1, 3^2, \ldots)) = 1 + \frac{(n-1)!}{(\beta^{-1}+n)^{n-1}} \sum_{i=0}^{n-2} \frac{\beta^{-1}(\beta^{-1}+i)^{i-1}(n-i)^{n-i-1}}{i!(n-i-2)!}$$

Proposition 3 *When $\boldsymbol{F}_J \sim \mathtt{BellP}_J(n, \theta_1/\lambda, \theta_2/\lambda, \ldots, \theta_J/\lambda, (1^0, 2^1, 3^2, \ldots))$,*

$$Cov(F_i, F_j | N = n) = -n\pi_i\pi_j\phi(n, \beta^{-1}, (1^0, 2^1, 3^2, \ldots)).$$

ベル多項式分布族による攪乱の安全性

- 攪乱に $\mathtt{BellP}_J(n, \lambda_1, \lambda_2, \ldots, \lambda_J, \boldsymbol{w})$ を使うとして差分プライバシーを評価する。
- 対数尤度を λ_j で微分すると

$$(\log B_{f_j}(\lambda_j \boldsymbol{w}) - \log B_n(\lambda_. \boldsymbol{w}))' = \frac{B'_{f_j}(\lambda_j \boldsymbol{w})}{B_{f_j}(\lambda_j \boldsymbol{w})} - \frac{B'_n(\lambda_. \boldsymbol{w})}{B_n(\lambda_. \boldsymbol{w})} \tag{10}$$

- これが全ての $(\boldsymbol{f}, (\lambda_1, \ldots, \lambda_J))$ について ϵ 以下なら差分プライベート。
- (10) 式の右辺第二項は $\boldsymbol{f}$ の変化について定数なので第一項のみ書き下す：

$$\frac{B'_{f_j}(\lambda_j \boldsymbol{w})}{B_{f_j}(\lambda_j \boldsymbol{w})} = \frac{\sum_{k=0}^{f_j} k\lambda_j^{k-1} B_{f_j,k}(\boldsymbol{w})}{\sum_{k=0}^{f_j} \lambda_j^k B_{f_j,k}(\boldsymbol{w})}$$

- 分母と分子を比較すると k/λ_j が大きいほど上式は増加する。
- 従って λ_j が最も小さい j に n を全て配分した場合に (10) 式は最大化される。

Theorem 2 *Suppose that* $n \in \mathbb{N}, J \in \mathbb{N}, \lambda_j > 0, w_j \geq 0, j \in [J]$. *Write* $\underline{\lambda} = \min_j \lambda_j$. *Then* $\texttt{BellP}_J(n, \lambda_1, \lambda_2, \ldots, \lambda_J, \boldsymbol{w})$ *is* ϵ*-DP if and only if*

$$\frac{B'_n(\underline{\lambda}\boldsymbol{w})}{B_n(\underline{\lambda}\boldsymbol{w})} - \frac{B'_n(\lambda.\boldsymbol{w})}{B_n(\lambda.\boldsymbol{w})} \leq \epsilon. \tag{11}$$

Corollary 2 *Suppose that* $n \in \mathbb{N}, J \in \mathbb{N}, p_j > 0, j \in [J]$, $\sum_{j=1}^{J} p_j = 1$. *Then* $\texttt{Multi}(n, \boldsymbol{p})$ *is* ϵ*-DP if and only if,* $\forall j$,

$$n(1/p_j - 1) \leq \epsilon. \tag{12}$$

- NB: $B_n(\lambda, 0, 0, \ldots) = \lambda^n$.
- When cell probabilites are uniform, i.e., least unsafe, (12) reduces to $n(J-1) \leq \epsilon$.
  - ϵ can never be small !

Corollary 3 *Suppose that* $n \in \mathbb{N}, J \in \mathbb{N}, \alpha_j > 0, j \in [J]$. *Then* $\texttt{NegHyp}(n, \vec{\alpha})$ *is* ϵ*-DP if and only if*

$$\sum_{k=0}^{n-1} \left(\frac{1}{\underline{\alpha} + k} - \frac{1}{\alpha. + k} \right) \leq \epsilon. \tag{13}$$

- When $\alpha./\underline{\alpha}$ is fixed at c, (13) reduces to

$$\sum_{k=0}^{n-1} \left(\frac{(c-1)\underline{\alpha}}{(\underline{\alpha} + k)(c\underline{\alpha} + k)} \right) \leq \epsilon.$$

  - LHS$\to 0$ as $\underline{\alpha} \to \infty$: $\texttt{Multi}$

Proposition 4 *Suppose that* $n \in \mathbb{N}, J \in \mathbb{N}, \alpha_j > 0, j \in [J]$. *Then for any positive* ϵ, $\texttt{NegHyp}(n, \vec{\alpha})$ *is* ϵ*-DP as* $\alpha. \to \infty$ *where cell probabilites are fixed.*

Remarks

- 負の超幾何（＝多項ディリクレ混合）分布が差分プライベートという結果は、セッティングが違うが Machanavajjhala et al. (2008) と整合的。
- 多項分布は母数を deterministic に攪乱しても一般に差分プライベートに出来ない。
  - 負の超幾何分布は多項分布の母数を混合し stochastic に攪乱している。
  - deterministic な攪乱の不確実性は尤度に出ない。

差分プライバシーの十分条件

- LHS of (11) is further written as

$$
\begin{aligned}
&\frac{\sum_{k=0}^{n} k\underline{\lambda}^{k-1} B_{n,k}(\boldsymbol{w}) B_n(\lambda_{\cdot}\boldsymbol{w}) - \sum_{k=0}^{n} k\lambda_{\cdot}^{k-1} B_{n,k}(\boldsymbol{w}) B_n(\underline{\lambda}\boldsymbol{w})}{B_n(\underline{\lambda}\boldsymbol{w}) B_n(\lambda_{\cdot}\boldsymbol{w})} \\
&= \frac{\sum_{k=0}^{n} k B_{n,k}(\boldsymbol{w}) \{\underline{\lambda}^{k-1} B_n(\lambda_{\cdot}\boldsymbol{w}) - \lambda_{\cdot}^{k-1} B_n(\underline{\lambda}\boldsymbol{w})\}}{B_n(\underline{\lambda}\boldsymbol{w}) B_n(\lambda_{\cdot}\boldsymbol{w})} \\
&\leq \frac{\sum_{k=0}^{n} k B_{n,k}(\boldsymbol{w}) \{\lambda_{\cdot}^{k-1} B_n(\lambda_{\cdot}\boldsymbol{w}) - \lambda_{\cdot}^{k-1} B_n(\underline{\lambda}\boldsymbol{w})\}}{B_n(\underline{\lambda}\boldsymbol{w}) B_n(\lambda_{\cdot}\boldsymbol{w})} \\
&\leq \frac{\sum_{k=0}^{n} n B_{n,k}(\boldsymbol{w}) \{\lambda_{\cdot}^{k-1} B_n(\lambda_{\cdot}\boldsymbol{w}) - \lambda_{\cdot}^{k-1} B_n(\underline{\lambda}\boldsymbol{w})\}}{B_n(\underline{\lambda}\boldsymbol{w}) B_n(\lambda_{\cdot}\boldsymbol{w})} \\
&= \frac{n}{\lambda_{\cdot}} \frac{B_n(\lambda_{\cdot}\boldsymbol{w}) \{B_n(\lambda_{\cdot}\boldsymbol{w}) - B_n(\underline{\lambda}\boldsymbol{w})\}}{B_n(\underline{\lambda}\boldsymbol{w}) B_n(\lambda_{\cdot}\boldsymbol{w})} = \frac{n}{\lambda_{\cdot}} \frac{\{B_n(\lambda_{\cdot}\boldsymbol{w}) - B_n(\underline{\lambda}\boldsymbol{w})\}}{B_n(\underline{\lambda}\boldsymbol{w})}.
\end{aligned}
$$

Proposition 5 *Suppose that the conditions of Theorem 2 hold. Then* $\texttt{BellP}_J(n, \lambda_1, \lambda_2, \ldots, \lambda_J, \boldsymbol{w})$ *is* ϵ*-DP if*

$$\frac{n}{\lambda.}\left(\frac{B_n(\lambda.\boldsymbol{w})}{B_n(\underline{\lambda}\boldsymbol{w})} - 1\right) \leq \epsilon.$$

- When $\lambda./\underline{\lambda}$ is fixed at c,

$$\lim_{\underline{\lambda}\to\infty} \frac{B_n(\lambda.\boldsymbol{w})}{B_n(\underline{\lambda}\boldsymbol{w})} - 1 = \lim_{\underline{\lambda}\to\infty} \frac{(c\underline{\lambda})^n B_{n,n}(\boldsymbol{w}) + O(\underline{\lambda}^{n-1})}{\underline{\lambda}^n B_{n,n}(\boldsymbol{w}) + O(\underline{\lambda}^{n-1})} - 1 = (c^n - 1)$$

Theorem 3 *Suppose that* $n \in \mathbb{N}, J \in \mathbb{N}, \lambda_j > 0, w_j \geq 0, j \in [J]$*. Then for any positive* ϵ*,* $\texttt{BellP}_J(n, \lambda_1, \lambda_2, \ldots, \lambda_J, \boldsymbol{w})$ *is* ϵ*-DP as* $\lambda. \to \infty$ *where cell probabilites are fixed.*

- When cell probabilites are fixed, `Multi` is unchanged as $\lambda. \to \infty$.

Proposition 6 *Let* $\rho_j = \theta_j/\lambda.$ *Then* $\texttt{QM}(n, \vec{\rho})$ *is* ϵ*-DP as* $\rho. \to \infty$ *where cell probabilities are fixed.*

- $B_n(\rho(1^0, 2^1, \ldots)) = \rho(\rho + n)^{n-1}$

まとめ

- 本報告では匿名性を確保した内訳情報の作成方法を考察した。
- 総額がトップコーディングで与えられていても、内訳の比は意味がある。
- 公表する比の下一桁を四捨五入するなら、総和が 1 になるとは限らない。
- random rounding なら総和は必ず 1 になるし安全性も上昇。
- random part がベル多項式分布族なら不偏性は成立する。
- しかし random part が多項分布では差分プライベートにならない。
- ベル多項式分布は多項分布に近いところで差分プライベートになる。
  - セル確率が小さい時に多くのボールが配られるという珍しいイベントの確率が、セル確率の変化について感応的なら差分プライベートでない。

多次元時系列トレンド・季節性・ノイズの SIML 分析と公的データへの応用

国友直人

明治大学政治経済学部

2017 年 12 月

佐藤整尚・栗栖大輔・粟屋直 (東京大学経済学研究科) との共同研究
Kunitomo-Sato(2017), Kunitomo-Awaya-Kurisu(2017) に基づく報告

Outline

問題の発端

2016 年-2017 年「速報性のある包括的な消費関連指標の在り方に関する研究会」(統計局) を巡る議論の考察から佐藤整尚・栗栖大輔・粟屋直 (東京大学大学院) の諸氏との共同研究へ発展。本日は主に研究動機、理論的考察 (Kunitomo-Sato(2017), Kunitomo-Awaya-Kurisu(2017), 未定稿)、今後の展開などを議論する
統計局・研究会の概要は既に報告書「消費動向指数 (CTI) に向けて」(統計局) で HP 上に公開、CTI は近い将来に実用化される予定。

問題の発端

時間の経過とともに多数の経済時系列が観察されるが、伝統的には多くのマクロ経済時系列データの場合にはデータの収集上・作成上の理由などから月次系列、四半期系列、年次系列などの離散時間単位で計測、公表。近年では計測法、情報処理の利便性が高まり、月次よりも高頻度の観測データも得られる経済データも存在。従来より高頻度の時系列データが利用可能になるにつれ、市場動向の理解も進むと考えられるが、他方、経済時系列ではマクロ経済の大局的な動きにも関心があり、月次、四半期、年次の動きとの整合性も重要。

具体的事例１：日次データの分解モデル

t は固定、月次周期 $m = 28, 29, 30, 31$ および四半期周期はともに変動することを考慮する必要。日次単位からは季節周期の構成日数は変動。時系列 y_t の加法的分解モデル

$$y_t = x_t + s_t^w + s_t^m + h_t + v_t \ \ (t = 1, \cdots, T),$$

を考察する。x_t はトレンド成分、s_t^w は週次成分、s_t^m は月次成分、h_t が特別休日成分、v_t は不規則変動である。加法的分解モデルを採用し、ま ｚ ｙ簡単化の為に循環成分をゼロとする。不規則成分 v_t は $N(0, \sigma^2)$ にしたがう互いに独立な確率変数、トレンド成分は

$$x_t = x_{t-1} + v_t^x \ \ (t = 1, \cdots, T),$$

にしたがい、不規則成分 v_t^x は確率分布 $N(0, \sigma_x^2)$ にしたがう互いに独立な確率変数、(Kitagawa (2010) を参考)

週次成分は

$$(1 + L + \cdots + L^6)s_t^w = v_t^w \ \ (t = 1, \cdots, T),$$

とする。L はラグ作用素、不規則成分 v_t^w は確率分布 $N(0, \sigma_w^2)$ にしたがう互いに独立な確率変数とする。月次成分は

$$(1 + L + \cdots + L^{11})[\sum_{t \in I_i(t)} s_t^m] = v_{I_i(t)}^m \ \ (t = 1, \cdots, T),$$

とする。月次時系列に基づく季節調整法とは整合的、月次状態成分は退化、結局 $s_{I_i(t)}^m = \sum_{t \in I_i(t)} s_i^m$ を推定すればよい。不規則成分 s_i^m を確率分布 $N(0, \sigma_m^2)$ にしたがう互いに独立な確率変数とすると、$v_{I_i(t)}^m$ の分散は σ_m^2 に比例する。
統計的問題として新しい観点は月次成分の扱いであり、状態空間表現では制約条件はかなり疎であるが、時間に依存、高次元問題となる。

消費の日次データ

利用するデータは 2000 年 1 月から 2016 年 10 月までの約 6000 の日次データである。一般にはあまり知られていないが、総務省統計局では 2000 年から日次データの集計を開始しているが、一般には月次データ、四半期データを様々な用途で基礎的に消費データとして利用されている。日次 2010.1.1 から 500 個のデータ、月次 2000.1-2016.12 のデータ。

具体的事例 2: 月次マクロ指標

マクロ経済データの場合には収集上・作成上の理由から日々の系列、月次系列、四半期系列、年次系列など様々な頻度と異なる時間的タイミングで計測され、公表。消費、投資、政府支出、輸出入など主要なマクロ時系列は調査や作成上の理由から相互に調整されて作成されていない。しかしマクロ経済の動向を理解、政策評価など行う立場からは望ましくない。また直近の状況を理解するためには早めのデータ作成が望ましいが、国全体のデータ作成には多くの情報が必要である。例えば GDP やその主要な項目は最速で四半期、数カ月後以降に公表される。後からより正確と思われるデータが利用可能となり、その結果として、公表した後に過去の公表数値が改訂されることも多い。直近に公表されている四半期データが直近で得られる月次系列などの情報と見かけ上で矛盾する事例もある。また日本の政府統計では担当部署が分かれていることで問題は複雑化。

マクロ消費動向は、サンプリングによる家計消費と商業動態統計など生産・販売の動向の乖離が重要な問題。家計調査データは世帯をベースにした標本調査の集計値、家計調査データでは世帯数の変化などの近年の動向を考慮して解釈する必要がある。他方、生産・販売の調査データには企業消費・政府消費やインバウンド消費など GDP における家計消費概念とは必ずしも整合的でない集計なので、マクロ消費を計測する際にはこうした項目の影響を勘案する必要がある。多くのエコノミストは GDP 最終消費の数値を重視、GDP 速報の推計では家計面と企業面における消費の情報を統合した数値を四半期ベースで作成、GDP 推計の確報は、生産面のより細かな推計値を主に利用 (内閣府 (2010))。

時系列分析の問題: 観察されるマクロ指標の四半期データ y_{1t} $(t = 3(i-1)+j\ ; i = 1, \cdots, m; j = 1, 2, 3;\ T = 3m)$. 関係するより高頻度な月次データを $p-1$ 次元ベクトル $\mathbf{y}_{2t}$ $(t = 12(i-1)+j, i = 1, \cdots, n; j = 1, \cdots, 12;\ T = 12n)$ とすると、利用可能な情報の下で観測不能な真のトレンドの状態 x_{1t} $(t = 12(i-1)+j, i = 1, \cdots, n; j = 1, \cdots, 12)$ を推定する問題。観測不能な真の非定常トレンド $\mathbf{x}_t = (x_{kt})$ 間に関係 $\beta'_{x,t}\mathbf{y}_t = O_p(1)$, また観測不能な真の非定常季節性 $\mathbf{s}_t = (s_{k,t})$ 間にも関係 $\beta'_{s,t}\mathbf{s}_t = O_p(1)$, が考えられる。
例えば Kunitomo-Sato(2017) はベクトル β_x,β_s を推定する SIML(分離情報最尤法) を提案。トレンド成分間の線形関係に注目し、季節成分ベクトル間の制約 $\beta'_x\mathbf{x}_t = \beta_0 + u_t^{(x)}$ 消費トレンド成分が共和分関係 (co-integrated relations) に対応。

消費系列への応用では最終消費系列 y_{1t} を構成する需要側時系列を y_{2t}、供給側時系列を y_{3t} として、線形関係 (あるいは共和分関係) を利用することで観測できない目的変数の月次系列のトレンドの状態推定の方法を考える。利用可能は四半期データと月次データの情報を有効に利用することで互いに矛盾のない状態推定を実現する必要がある。
この問題を解決するにはいくつかの統計的課題がある。
(i) 非定常多次元時系列でのフィルタリング問題,
(ii) 多次元時系列における季節性の処理 (i.e. 季節調整),
(iii) 非定常多次元時系列での母数推定,
(iv) 非定常多次元時系列における不完全観測におけるフィルタリング、
などが主な課題。逆にこうした問題について解決できれば様々な応用問題が解決できる。例えば現行の GDP 推計では季節調整系列の作成は整合的といえるか？などの問題 etc.

最適な状態推定の問題

1 次元時系列が完全観測の場合には統計的フィルタリング法として X-12-ARIMA, DECOMP などがあり、季節調整などに応用されている。多次元時系列の場合は完全観測の場合においても実用的な統計的フィルタリング法は存在せず。

非定常状態変数に関係 (あるいは構造方程式) があるとき $\mathrm{ranl}(\mathbf{x}_t\mathbf{x}_t^{'})$ が確率的に $p-1$ となり、時間的に変動するベクトル $\beta_{x,t}$ が存在して $\mathbf{x}_t^{'}\beta_{x,t} = O_p(1)$ となるので考慮する必要がある。統計的フィルタリングでは係数ベクトル $\beta_{x,t}$ に階数条件を用いて状態推定を行う必要がある。例えば完全観測の場合には尤度関数は初期条件 $\mathbf{Y}_0$ の下、母数 $\boldsymbol{\theta}$ に対し

$$L_T(\boldsymbol{\theta}) = \prod_{t=1}^{T} f(\mathbf{y}_t|\mathbf{Y}_{t-1}, \boldsymbol{\theta})$$

を利用する必要がある。(例えば DECOMP では hyper-parameter と呼んでいるが) 特に非定常多次元の場合には推定などは自明でない。

SIML

マクロ SIML 法のアイデアを簡単な設定で説明:

y_{ij}: 第 j 変数の第 i 観測値
(時刻 t_i^n; $i = 1, \cdots, n; j = 1, \cdots, p$; $0 = t_0^n \leq \cdots \leq t_n^n = T$)
しばしば $n = T$, $t_i^n - t_{i-1}^n = 1$, $\mathbf{y}_i = (y_{1i}, \cdots, y_{pi})^{'}$: $p \times 1$ ベクトル, $\mathbf{Y}_n = (\mathbf{y}_i^{'})$ $(= (y_{ij}))$: $n \times p$ 観測行列, $\mathbf{y}_0$: 初期ベクトル.
非定常トレンド $\mathbf{x}_i$ $(= (x_{ji}))$ at t_i^n $(i = 1, \cdots, n)$ は観測ベクトルと同一ではない。季節要素 $\mathbf{s}_i^{'} = (s_{1i}, \cdots, s_{pi})$. ノイズ要素 $\mathbf{v}_i^{'} = (v_{1i}, \cdots, v_{pi})$ とするが、簡単化のためにトレンドとは独立な系列と仮定.

加法分解モデル

$$\mathbf{y}_i = \mathbf{x}_i + \mathbf{s}_i + \mathbf{v}_i$$

ここで $\mathbf{x}_i$ 非定常トレンド要素, $\mathbf{s}_i$ 非定常季節要素, ノイズ要素は定常確率過程であり

$$\Delta^d \mathbf{x}_i = (1 - \mathcal{L})^d \mathbf{x}_i = \mathbf{w}_i^{(x)}$$

を満たす。ただし $\mathcal{L}\mathbf{x}_i = \mathbf{x}_{i-1}$, $\Delta = 1 - \mathcal{L}$, $\mathcal{E}(\mathbf{w}_i^{(x)}) = \mathbf{0}$, $\mathcal{E}(\mathbf{w}_i^{(x)} \mathbf{w}_i^{(x)'}) = \mathbf{\Sigma}_x$,

$$(1 + \mathcal{L} + \cdots + \mathcal{L}^{s-1})^D \mathbf{s}_i = \mathbf{w}_i^{(s)}$$

$\mathcal{L}^s \mathbf{s}_i = \mathbf{s}_{i-s}$, $\mathcal{E}(\mathbf{w}_i^{(s)}) = \mathbf{0}$, $\mathcal{E}(\mathbf{w}_i^{(s)} \mathbf{w}_i^{(s)'}) = \mathbf{\Sigma}_s$,

$$\mathbf{v}_i = \sum_{j=-\infty}^{\infty} \mathbf{C}_j \mathbf{e}_{i-j} \ ,$$

ここで係数行列 $\mathbf{C}_j$ の要素の絶対値和は収束, i.i.d. 確率ベクトルは $\mathcal{E}(\mathbf{e}_i) = \mathbf{0}$, $\mathcal{E}(\mathbf{e}_i \mathbf{e}_i^{'}) = \mathbf{\Sigma}_e$ を満たす。

簡単化のためにここでは $d = D = 1$ とおき、観測されない非定常確率変数の間には構造関係が存在する状況を考察すると、非定常変数誤差モデルになる。例えば $\boldsymbol{\beta}$ $p \times 1$ ベクトルとして構造関係 $\boldsymbol{\beta}^{'}\mathbf{y}_i = O_p(1)$ $(i = 1, \cdots, n)$, あるいはよりっぱん的に $\mathbf{B}$ $q \times p$ $(q \leq p)$ 行列とすると、

$$\mathbf{B}\mathbf{y}_i = O_p(1) \ \ (i = 1, \cdots, n)$$

の推定問題を考察する。同様に季節要素間の関係

$$\mathbf{B}_s\mathbf{s}_i = O_p(1) \ \ (i = 1, \cdots, n).$$

を分析する必要がある。
この問題はトレンド・ベクトルや季節性・ベクトルの階数が退化する問題、したがって既存の統計的方法を拡張する必要がある。

ここでは議論を単純化して $d = D = 1$, 季節性がない場合 ($\mathbf{s}_i = 0$) を考察する。さらに $\Delta\mathbf{x}_i$ と $\mathbf{v}_i$ $(i = 1, \cdots, n)$ は独立、各要素はガウス分布 $N_p(\mathbf{0}, \boldsymbol{\Sigma}_x)$ および $N_p(\mathbf{0}, \boldsymbol{\Sigma}_v)$ にしたがうとする。記号は $n \times p$ 行列 $\mathbf{Y}_n = (\mathbf{y}_i^{'})$, であり $np \times 1$ 確率ベクトル $(\mathbf{y}_1^{'}, \cdots, \mathbf{y}_n^{'})^{'}$ の分布は初期ベクトル $\mathbf{y}_0$ の下で

$$\mathrm{vec}(\mathbf{Y}_n) \sim N_{n \times p}\left(\mathbf{1}_n \cdot \mathbf{y}_0^{'}, \mathbf{I}_n \otimes \boldsymbol{\Sigma}_v + \mathbf{C}_n\mathbf{C}_n^{'} \otimes \boldsymbol{\Sigma}_x\right) \ ,$$

ここで行列

$$\mathbf{C}_n = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 0 \\ 1 & \cdots & 1 & 1 & 0 \\ 1 & \cdots & 1 & 1 & 1 \end{pmatrix}_{n \times n} .$$

とすると、初期条件 $\mathbf{y}_0$ の下で条件付最尤 (ML) 推定量は条件付対数尤度関数

$$\begin{aligned} L_n^* &= \log |\mathbf{I}_n \otimes \mathbf{\Sigma}_v + \mathbf{C}_n \mathbf{C}_n^{'} \otimes \mathbf{\Sigma}_x|^{-1/2} \\ &- \frac{1}{2}[vec(\mathbf{Y}_n - \bar{\mathbf{Y}}_0)^{'}]^{'}[\mathbf{I}_n \otimes \mathbf{\Sigma}_v + \mathbf{C}_n \mathbf{C}_n^{'} \otimes \mathbf{\Sigma}_x]^{-1}[vec(\mathbf{Y}_n - \bar{\mathbf{Y}}_0)^{'}] \quad , \end{aligned}$$

の最大化問題であるが、

$$\bar{\mathbf{Y}}_0 = \mathbf{1}_n \cdot \mathbf{y}_0^{'} .$$

ここで K_n- 変換（$\mathbf{Y}_n$ より $\mathbf{Z}_n\,(=(\mathbf{z}_k^{'}))$）を次のようにとる

$$\mathbf{Z}_n = \mathbf{K}_n \left(\mathbf{Y}_n - \bar{\mathbf{Y}}_0\right) \, , \mathbf{K}_n = \mathbf{P}_n \mathbf{C}_n^{-1} \, ,$$

ただし

$$\mathbf{C}_n^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & 0 & \cdots \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}_{n \times n} ,$$

および

$$\mathbf{P}_n = (p_{jk}^{(n)}) \, , \; p_{jk}^{(n)} = \sqrt{\frac{2}{n+\frac{1}{2}}} \cos\left[\frac{2\pi}{2n+1}(k-\frac{1}{2})(j-\frac{1}{2})\right] .$$

固有値分解より $\mathbf{C}_n^{-1}\mathbf{C}_n^{\prime -1} = \mathbf{P}_n\mathbf{D}_n\mathbf{P}_n^{\prime}$ ただし $\mathbf{D}_n$ が対角行列であり k-要素は

$$d_k = a_{kn}^* = 2[1-\cos(\pi(\frac{2k-1}{2n+1}))] = 4\sin^2(\pi/2)[(2k-1)/(2n+1)]$$

初期値を所与とする条件付対数尤度関数は定数を除き

$$L_n = \sum_{k=1}^{n} \log|a_{kn}\mathbf{\Sigma}_v + \mathbf{\Sigma}_x|^{-1/2} - \frac{1}{2}\sum_{k=1}^{n} \mathbf{z}_k^{'}[a_{kn}\mathbf{\Sigma}_v + \mathbf{\Sigma}_x]^{-1}\mathbf{z}_k \ ,$$

ここで

$$a_{kn}\ (=d_k)\ = 4\sin^2\left[\frac{\pi}{2}\left(\frac{2k-1}{2n+1}\right)\right]\ (k=1,\cdots,n)\ .$$

変数変換することにより非定常多次元時系列 $\mathbf{z}_k\ (k=1,\cdots,n)$ は $N_p(\mathbf{0}, \mathbf{\Sigma}_x + a_{kn}\mathbf{\Sigma}_v)$, にしたがうが、係数 a_k は関数 $4\sin^2(x)$ in $(0,\pi/2)$ の離散値をとる

$\mathbf{z}_k$ の分散共分散行列 $a_{kn}\mathbf{\Sigma}_v + \mathbf{\Sigma}_x$ を推定するには $\mathbf{z}_k\mathbf{z}_k^{'}$ を利用するのが自然であるが、k を固定すると $n \to \infty$ のとき $a_{kn} \to 0$ となる。k が小さければ a_{kn} は小さい。$k = k_n$ とする。このとき $(1/m_n)\sum_{k=1}^{m_n} a_{kn}$ は m_n が n に近いと小さくないが $m_n/n \to 0 (n \to \infty)$ である。分離最尤推定 (SIML) 量は $(\hat{\mathbf{\Sigma}}_x)$

$$\hat{\mathbf{\Sigma}}_{x,SIML} = \frac{1}{m_n}\sum_{k=1}^{m_n} \mathbf{z}_k\mathbf{z}_k^{'} .$$

周波数領域で考えると、観測誤差が存在する場合の非定常トレンドの分散共分散行列の推定量となっている。$\hat{\mathbf{\Sigma}}_x$ に対して項数 m_n は n に依存するようにとり、$m_n = O(n^{\alpha})(0 < \alpha < 1)$ とするのがマクロ SIML 推定量である。

非定常共通 1 トレンド・モデルの場合

ここでトレンド＋ノイズの多次元非定常時系列においてトレンド階数が 1 次元に退化する場合を考察する。
共和分ベクトルが一つ存在する場合を考察する。$(q = 1)$ ここで $\mathbf{y}_i = \mathbf{x}_i + \mathbf{v}_i$, $\mathbf{Y}_n = (\mathbf{y}_i^{'})$, $\mathbf{x}_i$ は

$$\mathbf{x}_i = \mathbf{x}_{i-1} + \boldsymbol{\pi}\mu_i ,$$

満足するとする。ここで $\boldsymbol{\pi}$ は $p \times 1$ vector, μ_i は i.i.d. (1 次元) 確率変数で $N(0, \sigma_{\mu}^2)$ にしたがい、$\mathbf{v}_i$ は i.i.d. (p-次元) 確率変数で $N_p(\mathbf{0}, \mathbf{\Sigma}_v)$ にしたがうことを仮定する。($\mathbf{\Sigma}_v$ は正則行列とする。)

さらに $\mathbf{b} = \sigma_\mu \boldsymbol{\pi}, \mathbf{A} = a_{kn}\boldsymbol{\Sigma}_v$ および逆行列の公式（正定符号行列 $\mathbf{A}$）

$$|\mathbf{A} + \mathbf{b}\mathbf{b}'| = |\mathbf{A}|[1 + \mathbf{b}'\mathbf{A}^{-1}\mathbf{b}]$$

$$[\mathbf{A} + \mathbf{b}\mathbf{b}']^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{b}[1 + \mathbf{b}'\mathbf{A}^{-1}\mathbf{b}]^{-1}\mathbf{b}'\mathbf{A}^{-1}$$

を利用する。$(\boldsymbol{\Sigma}_x = \mathbf{b}\mathbf{b}')$

対数尤度関数 L_n は

$$\begin{aligned} L_{1n} &= \sum_{k=1}^{n} \left[\log |a_{kn}\boldsymbol{\Sigma}_v| + \log(1 + a_{kn}^{-1}\mathbf{b}'\boldsymbol{\Sigma}_v^{-1}\mathbf{b}) + a_{kn}^{-1}\mathbf{z}_k'\boldsymbol{\Sigma}_v^{-1}\mathbf{z}_k \right. \\ &\qquad \left. - \frac{a_{kn}^{-1}(\mathbf{z}_k'\boldsymbol{\Sigma}_v^{-1}\mathbf{b})^2}{a_{kn} + \mathbf{b}'\boldsymbol{\Sigma}_v^{-1}\mathbf{b}} \right] \\ &= \sum_{k=1}^{n} \log |a_{kn}\boldsymbol{\Sigma}_v| + \sum_{k=1}^{n} a_{kn}^{-1}\mathbf{z}_k'\boldsymbol{\Sigma}_v^{-1}\mathbf{z}_k \\ &\qquad + \sum_{k=1}^{n} \left[\log(1 + a_{kn}^{-1}c) - \frac{a_{kn}^{-1}(\mathbf{z}_k'\boldsymbol{\Sigma}_v^{-1}\mathbf{b})^2}{a_{kn} + c} \right] , \end{aligned}$$

に反比例する。ここで

$$c = \sigma_\mu^2 \boldsymbol{\pi}'\boldsymbol{\Sigma}_v^{-1}\boldsymbol{\pi} \ .$$

ただしベクトル $\boldsymbol{\pi}$ には基準化が必要、ML 推定量 $\boldsymbol{\pi}$ は二次元でも複雑なかいとなるが、ここでは $\boldsymbol{\beta}^{'} = (1, -\boldsymbol{\beta}_2^{'})$ としておく。

$$\left[\hat{\boldsymbol{\Sigma}}_{x.SIML} - \lambda \hat{\boldsymbol{\Sigma}}_{v.SIML}\right] \hat{\boldsymbol{\beta}}_{SIML} = \mathbf{0} \ ,$$

$$\hat{\Sigma}_{x.SIML} = \frac{1}{m_n} \sum_{k=1}^{m_n} \mathbf{z}_k \mathbf{z}_k^{'} \ ,$$

$$\hat{\Sigma}_{v.SIML}(1) = \frac{1}{2}[\frac{1}{n} \sum_{k=1}^{n} \mathbf{z}_k \mathbf{z}_k^{'} - \hat{\Sigma}_{x.SIML}] \ ,$$

あるいは

$$\hat{\Sigma}_{v.SIML}(2) = \frac{1}{l_n} \sum_{k=n+1-l_n}^{n} a_{kn}^{-1} \mathbf{z}_k \mathbf{z}_k^{'} - \frac{1}{4} \hat{\Sigma}_{x.SIML} \ ,$$

ただし

$$\mathbf{Z}_n = (\mathbf{z}_k^{'}) = \mathbf{P}_n \mathbf{C}_n^{-1} \left(\mathbf{Y}_n - \mathbf{1}_n \bar{\mathbf{y}}_0^{'} \right) \ ,$$

$\hat{\mathbf{\Sigma}}_{v.SIML}$ を SIML 推定量 ($\mathbf{\Sigma}_v$) と呼ぶが、λ は固有値である。

行列 $\mathbf{\Sigma}_x$ の階数は退化して 1 であるから最小固有値 λ_1 をとると、$\hat{\boldsymbol{\beta}}_{SIML}$ は $\boldsymbol{\beta}$ の SIML 推定量となる。より簡単な推定量は

$$\hat{\mathbf{\Sigma}}_{x.SIML} \times \hat{\boldsymbol{\beta}}_{SIL} = \mathbf{0} \ ,$$

すなわち

$$\hat{\mathbf{\Sigma}}_{x.SIML} \times \begin{bmatrix} 1 \\ -\hat{\boldsymbol{\beta}}_{2.SIL} \end{bmatrix} = \mathbf{0} \ .$$

解けば

$$\hat{\boldsymbol{\beta}}_{2.SIL} = \hat{\mathbf{\Sigma}}_{22x.SIML}^{-1} \hat{\mathbf{\Sigma}}_{21x.SIML} \ ,$$

となり一種の最小二乗推定量となる。

ガウス尤度

トレンド＋ノイズの二次元非定常時系列の尤度関数 (二次元 DECOMP モデル) の挙動を考察

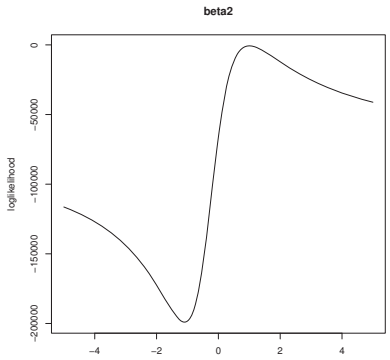

Figure 1 : Likelihood Function of β_2 $(n = 1,000)$

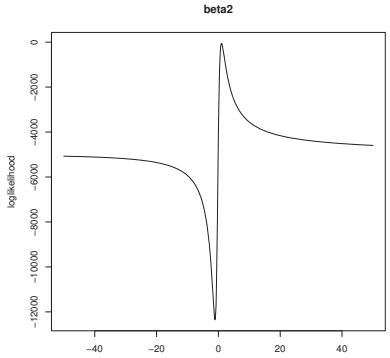

Figure 2 : Likelihood Function of β_2 $(n = 1,000)$

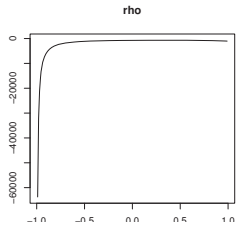

Figure 3 : Likelihood Function of ρ ($n = 1,000$)

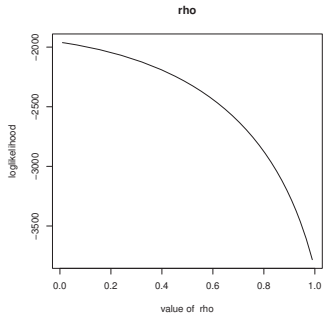

Figure 4 : Wrong Likelihood Function of ρ ($n = 1,000$)

二次元非定常モデルの最尤推定と SIML 推定の漸近的性質

Theorem 4.2 (KAK 2017)) : Assume that $\mathbf{v}_i$ ($i = 1, \cdots, n$) are a squence of i.i.d. random vectors and $|\mathbf{\Sigma}_v| \neq 0$. Then under the assumption of Gaussian distributions the maximum likelihood estimator of $\boldsymbol{\beta}$ is consistent as $n \longrightarrow \infty$.

Theorem 4.3 (KS 2017) : Assume the non-stationary errors-in-variables model and $|\mathbf{\Sigma}_v| \neq 0$. Then under the assumption of existence of fourth order moments the SIML estimator of $\boldsymbol{\beta}$ is consistent as $n \longrightarrow \infty$.

なお SIML 推定量の漸近正規性も示せるが省略する (KS(2017).)

(ノイズに自己相関が存在する場合)
ノイズのスペクトル行列が

$$f_v(\lambda) = \frac{1}{\pi}(\sum_{j=-\infty}^{\infty} \mathbf{C}_j e^{2i\lambda j})\mathbf{\Sigma}_e(\sum_{j=-\infty}^{\infty} \mathbf{C}_j e^{-2i\lambda j}) \quad (-\frac{\pi}{2} \leq \lambda \leq \frac{\pi}{2}),$$

のとき変換された確率過程のスペクトル密度行列
$(\Delta \mathbf{y}_i\ (= \mathbf{y}_i - \mathbf{y}_{i-1}))$ は

$$f_{\Delta y}(\lambda) = \frac{1}{\pi}\left[\mathbf{\Sigma}_x + (1 - e^{2i\lambda}) f_v(\lambda)(1 - e^{-2i\lambda})\right] \quad .$$

季節成分の推定

多次元非定常時系列における季節性の SIML 推定

季節性を含む場合には $\mathbf{y}_i = \mathbf{x}_i + \mathbf{s}_i + \mathbf{v}_i\ (i = 1, \cdots, n)$, ただし $\mathbf{x}_i$ トレンド成分, $\mathbf{s}_i$ 季節成分, $\mathbf{v}_i$ 観測誤差成分とする。階差作用素 $\Delta = 1 - \mathcal{L}\ (\mathcal{L}\mathbf{y}_i = \mathbf{y}_{i-1})$ および変換

$$\mathbf{B}_n^{(3)} = (b_{jk}^{(3)}) = \mathbf{P}_n \mathbf{C}_n^{-2} \mathbf{C}_n^{(s)} \ ,$$

を利用する。ここで $\mathbf{C}_n^{(s)} = \mathbf{C}_N \otimes \mathbf{I}_s$, ただし $N, s\ (\geq 2)$, $n = Ns$ 正整数とする。このとき

$$\sum_{j=1}^{n} b_{kj}^{(3)} b_{k',j}^{(3)} = 4\delta(k, k') \frac{\sin^4\left[\frac{\pi}{2}\frac{2k-1}{2n+1}\right]}{\sin^2\left[\frac{\pi}{2}\frac{2k-1}{2n+1}s\right]} + o(\frac{1}{n}) \ .$$

小さな相関を無視 $O(n^{-1})$ できるので、基準関数は

$$\begin{aligned} L_n^{(SI)} &= \sum_{k=1}^{n} \log |a_{kn}^* \mathbf{\Sigma}_v + a_{kn}^{(s)} \mathbf{\Sigma}_s + \mathbf{\Sigma}_x|^{-1/2} \\ &\quad -\frac{1}{2} \sum_{k=1}^{n} \mathbf{z}_k' [a_{kn}^* \mathbf{\Sigma}_v + a_{kn}^{(s)} \mathbf{\Sigma}_s + \mathbf{\Sigma}_x]^{-1} \mathbf{z}_k \ , \end{aligned}$$

ただし a_{kn}^* は

$$a_{kn}^{(s)} = 4 \frac{\sin^4 \left[\frac{\pi}{2} \left(\frac{2k-1}{2n+1} \right) \right]}{\sin^2 \left[\frac{\pi}{2} \left(\frac{2k-1}{2n+1} s \right) \right]} \ (k = 1, \cdots, n) \ .$$

季節性を含むとき次の結果が得られる。

Theorem 5.1 : In the new setting with $N, s, n\ (= Ns)$ (positive integers), we assume the moment conditions on the seasonal components as $\mathcal{E}[w_{ig}^{(s)4}] < \infty$.
For $\hat{\mathbf{\Sigma}}_x$, we have
(i) For $m_n = [n^\alpha]$ and $0 < \alpha < 1$, as $n \longrightarrow \infty$

$$\hat{\mathbf{\Sigma}}_x - \mathbf{\Sigma}_x \xrightarrow{p} \mathbf{O} \ .$$

(ii) For $m_n = [n^\alpha]$ and $0 < \alpha < 0.8$, as $n \longrightarrow \infty$

$$\sqrt{m_n} \left[\hat{\sigma}_{gh}^{(x)} - \sigma_{gh}^{(x)} \right] \xrightarrow{\mathcal{L}} N \left(0, \sigma_{gg}^{(x)} \sigma_{hh}^{(x)} + \left[\sigma_{gh}^{(x)} \right]^2 \right) \ .$$

The covariance of the limiting distributions of $\sqrt{m_n}[\hat{\sigma}_{gh}^{(x)} - \sigma_{gh}^{(x)}]$ and $\sqrt{m_n}[\hat{\sigma}_{kl}^{(x)} - \sigma_{kl}^{(x)}]$ is given by
$\sigma_{gk}^{(x)} \sigma_{hl}^{(x)} + \sigma_{gl}^{(x)} \sigma_{hk}^{(x)}$ $(g, h, k, l = 1, \cdots, p)$.

季節要素の分散共分散行列 $\mathbf{\Sigma}_s = (\sigma_{gh}^{(s)})$ は推定量 $\hat{\mathbf{\Sigma}}_s = (\hat{\sigma}_{gh}^{(s)})$, を用いるが、ここで

$$\hat{\mathbf{\Sigma}}_{s,SIML} = \frac{1}{m_n} \sum_{k \in \mathrm{I}_n^{(s)}} a_{kn}^{(s)-1} \mathbf{z}_k \mathbf{z}_k^{'} \ ,$$

ただし s 季節整数, $[x]$ ガウス記号, $\mathrm{I}_n^{(s)}$ は集合
$\mathrm{I}_{1n}^{(s)} = \{[2n/s]+1, \cdots, [2n/s]+m_n]\}$ ただし
$m_n = [n^{\alpha}]$ $(0 < \alpha < 1)$. あるいは
$\mathrm{I}_{2n}^{(s)} = \{[2n/s]-[m_n/2], \cdots, [2n/s], \cdots, [2n/s]+[m_n/2]]\}$ を用いても良い。
ここで $[2n/s]$ は季節数端数に対応する。($s=4$, $s=12$.)

ここでトレンド・季節性・観測誤差間には

$$\mathcal{E}[\mathbf{z}_k \mathbf{z}_k^{'}] = \mathbf{\Sigma}_x + a_{kn}^{(s)} \mathbf{\Sigma}_s + a_{kn}^{*} \mathbf{\Sigma}_v \ .$$

したがって

$$\mathcal{E}[a_{kn}^{(s)-1} \mathbf{z}_i \mathbf{z}_i^{'}] = \mathbf{\Sigma}_s + a_{kn}^{(s)-1} \mathbf{\Sigma}_x + \frac{a_{kn}^{*}}{a_{kn}^{(s)}} \mathbf{\Sigma}_v \ .$$

季節要素の分散共分散行列の推定について次の結果が成り立つ。

Theorem 5.2 : In the new setting assume the moment conditions on the seasonal components as $\mathcal{E}[w_{ig}^{(s)4}] < \infty$. For $\hat{\mathbf{\Sigma}}_s$ with $\mathrm{I}_{1n}^{(s)}$ or $\mathrm{I}_{2n}^{(s)}$,
(i) for $m_n = [n^\alpha]$ and $0 < \alpha < 1$, as $n \longrightarrow \infty$

$$\hat{\mathbf{\Sigma}}_s - \mathbf{\Sigma}_s \xrightarrow{p} \mathbf{O} \, .$$

(ii) For $m_n = [n^\alpha]$ and $0 < \alpha < 0.8$, as $n \longrightarrow \infty$

$$\sqrt{m_n}\left[\hat{\sigma}_{gh}^{(s)} - \sigma_{gh}^{(s)}\right] \xrightarrow{\mathcal{L}} N\left(0, \sigma_{gg}^{(s)}\sigma_{hh}^{(s)} + \left[\sigma_{gh}^{(s)}\right]^2\right) \, .$$

The covariance of the limiting distributions of $\sqrt{m_n}[\hat{\sigma}_{gh}^{(s)} - \sigma_{gh}^{(s)}]$ and $\sqrt{m_n}[\hat{\sigma}_{kl}^{(s)} - \sigma_{kl}^{(s)}]$ is given by $\sigma_{gk}^{(s)}\sigma_{hl}^{(s)} + \sigma_{gl}^{(s)}\sigma_{hk}^{(s)}$ $(g, h, k, l = 1, \cdots, p)$.

まとめと課題

- ファイナンス・SIML はマイクロノイズを含む高頻度金融データから integrated-volatility, co-volatility, Quadratic Variation の推定に有効
- マクロ・SIML はノイズ要素を含む場合に非定常トレンド・季節性における変数間の関係を推定する際に有効
- ノイズを含むとき非定常トレンド・季節性の共通ファクター数を推定することが重要
- 非定常トレンド・季節性の階数を推定する場合に SIML 推定で用いる分散共分散を利用することが自然、固有値・固有ベクトルを利用することを開発中
- 多次元非定常時系列におけるフィルタリングの実用化を北川源四郎氏・佐藤整尚氏などと開発中

文献

- Kitagawa, G. (2010), *Introduction to Time Series Modeling*, CRC Press.
- Anderson, T.W. (1984), "Estimating Linear Statistical Relationships," *Annals of Statistics*, 12, 1-45.
- Johansen, S. (1995), *Likelihood Based Inference in Cointegrated Vector Autoregressive Models*, Oxford UP.
- Kunitomo, N, S. Sato, and D. Kurisu (2017), *Separating Information Maximum Likelihood Method for High-frequency Financial Data*, Springer, in preparation.
- Kunitomo, N, and S. (2008), "Separating Information Maximum Likelihood Estimation in of Realized Volatility and Covariance with Micro-Market Noise," Discussion Paper CIRJE-F-581, (http://www.e.u-tokyo.ac.jp/cirje/research/dp/2008), also in *North American Journal of Economics and Finance* (2013).
- Kunitomo, N. and S. Sato (2017), "Trend, Seasonality and Economic Time Series : the Non-stationary Errors-in-variables Models," DP(SDS-4), MIMS, Meiji University, (http://www.mims.meiji.ac.jp/publications/datascience.html)
- Kunitomo, N., N. Awaya and D. Kurisu (2017), "Some Properties of Estimation Methods for Structural Relationships in Non-stationary Errors-in-Variables Models," DP(SDS-3), MIMS, Meiji University.