ランダムネスによる プライバシー保護の評価

星野　伸明

金沢大学・経済学経営学系

2019 年 1 月 31 日

本研究は科研費、統計数理研究所共同利用研究経費の助成を受けている。

概要

1. ランダムな個票データの情報漏洩危険性
2. 差分プライバシーと微分プライバシー
3. 各種ランダム度数メカニズムの危険性評価
4. 有望なメカニズム — 擬似多項分布の紹介

ランダムな個票データは安全？

- 情報保護のため、公開する個票データを乱数にすることはよく行われる。
  - 例）サンプリング、ランダムノイズ付加、ランダムスワッピング、模造
- 公開個票データから母集団の母数が推定されると情報漏洩が判明：
  - 既知属性所与で条件付き（母集団）分布のサイズが 1 と分かれば、識別開示
  - 既知属性所与で条件付き（母集団）分布が退化と分かれば、属性開示
  - 既知属性所与で条件付き（母集団）分布の分散が小と分かれば、推測開示
- ランダムな出力（＝標本）から母数（＝母集団分布）を推定する際の精度を、情報漏洩の危険性と考える。
- 母数推定精度（推定量の分散）は、対数尤度関数の傾きが決めている。
- 「差分プライバシー」は対数尤度関数の傾きを押さえることで、母数推定の精度を押さえている。

差分プライバシー) Machanavajjhala (2008) の枠組み

- $\vec{m} = (m_1, m_2, \ldots, m_J)$ が「標本」。m_j は第 j セルの観測度数。
  - $m = \sum_{j=1}^{J} m_j$ は標本サイズ。
- $\vec{n} = (n_1, n_2, \ldots, n_J)$ が「母集団」。n_j は第 j セルの母集団度数。
  - $n = \sum_{j=1}^{J} n_j$ は母集団サイズ。
- $\vec{n'} = (n'_1, n'_2, \ldots, n'_J)$ は $\vec{n}$ と一個体のみ違う度数ベクトル。つまり適当な $j \neq k$ について、$n'_j = n_j + 1, n'_k = n_k - 1$ となっている。
- ϵ-差分プライベート $\Leftrightarrow \forall(\vec{m}, \vec{n})$

$$\left| \mathsf{P}(\vec{m}; \vec{n}) / \mathsf{P}(\vec{m}; \vec{n'}) \right| \leq \exp(\epsilon) \tag{1}$$

Notes

- $\vec{n}$ と $\vec{n'}$ の差は一個体の移動だが、追加（ないし削除）した場合の尤度比で差分プライバシーを定義してもよいはず。
  - 移動では母集団に含まれていることに変わりはない。
- 「標本」を個票データとして公開するのでなければ、$\vec{m}$ が非負整数のベクトルである必要はない。
  - $\vec{n}$ の各要素にラプラースノイズを乗せた $\vec{m}$ は差分プライベートである。
  - 差分プライベートな実数のベクトルを事後処理して非負整数のベクトル化しても差分プライベート。しかしこの方法の挙動解析は面倒。
    - 寺田ほか (2016) は総和条件を満たす非負整数となるように事後処理。
  - 本報告では総和条件を満たす非負整数の多変量分布を考察。

微分プライバシー

- 一個体を追加ないし削除した場合の尤度比で差分プライバシーを定義：

$$\left| \frac{\log \mathsf{P}(\vec{m};\vec{n}) - \log \mathsf{P}(\vec{m};\vec{n}+\Delta)}{|\Delta|} \right| \leq \epsilon, \quad |\Delta| = 1$$

- ここで $|\Delta| \to 0$ の極限の解釈：

$$|\partial \log \mathsf{P}(\vec{m};\vec{n})/\partial n_j| \leq \epsilon \tag{2}$$

- 全ての $\vec{m}$ について (2) 式が成立するなら、正則条件の下で

$$\mathsf{V}(\hat{n_j}) \geq 1/\epsilon^2$$

  - $I(n_j) := \mathsf{E}_{\vec{m}}[(\partial \log \mathsf{P}(\vec{m};\vec{n})/\partial n_j)^2] \leq \epsilon^2$ である。この時クラメールラオの不等式が成立するなら $\mathsf{V}(\hat{n_j}) \geq 1/I(n_j)$
- つまり母数推定精度の上限を与えている。

考察）単純無作為非復元抽出

- リサンプリング機構として単純無作為非復元抽出を採用した場合（超幾何分布）：

$$\mathsf{P}(\vec{m};\vec{n}) = \frac{\binom{n_1}{m_1}\binom{n_2}{m_2}\cdots\binom{n_J}{m_J}}{\binom{n}{m}}$$

- この時 ϵ-差分プライバシーは達成出来ない。
  - $\mathsf{P}(\vec{m};\vec{n})/\mathsf{P}(\vec{m};\vec{n'})$ の分母が 0、分子が正となる場合がある。ex) $n_k = 1, m_k = 1$
- $\vec{m}$ の値域が母数 $\vec{n}$ に依存しているのが原因。
- $\vec{m}$ の値域を $\vec{n}$ に依存しないようにするには、例えば

$$\mathsf{P}(\vec{m};\vec{n},\vec{l}) = \frac{\binom{n_1+l_1}{m_1}\binom{n_2+l_2}{m_2}\cdots\binom{n_J+l_J}{m_J}}{\binom{n+l.}{m}} \tag{3}$$

  のような方法が考えられる。これは一種の模造機構である。
  - $l_j > m$ の時に $\vec{m}$ の値域は $\vec{n}$ に依存しない。

考察）超幾何分布による模造

- 模造機構として (3) 式を採用。変形した結果：

$$\mathsf{P}(\vec{m};\vec{n},\vec{l}) = \binom{m}{\vec{m}}\frac{\Gamma(n-m+l.+1)}{\Gamma(n+l.+1)}\prod_{j=1}^{J}\frac{\Gamma(n_j+l_j+1)}{\Gamma(n_j+l_j-m_j+1)}$$

- この時 ϵ-差分プライベート $\Leftrightarrow$

$$\left|\frac{1+\min_j l_j}{1+\min_j l_j - m}\right| \leq \exp(\epsilon) \tag{4}$$

  - 注）$\forall j,\ l_j > m$
- $n \gg m$ なら実用性があるかもしれないが。
- Machanavajjhala(2008) の例：$(m = \text{百万}\ , \epsilon = 7) \Rightarrow l_j \geq 1000911.714$

超幾何分布の微分プライバシー

- 微分プライバシー $\Leftrightarrow \forall(\vec{m},\vec{n})$

$$|\frac{\partial \log \mathsf{P}(\vec{m};\vec{n})}{\partial n_j}| = |\frac{1}{n_j+l_j}+\frac{1}{n_j+l_j-1}+\cdots+\frac{1}{n_j+l_j-m_j+1} - \frac{1}{n+l.}-\frac{1}{n+l.-1}-\cdots-\frac{1}{n+l.-m+1}| \leq \epsilon$$

- 上式左辺が最大化されるのは $n_j=0, m_j=m$ の時なので $(l_j > m)$、微分プライバシー $\Leftrightarrow$

$$|\frac{1}{\min_j l_j}+\frac{1}{\min_j l_j-1}+\cdots+\frac{1}{\min_j l_j-m+1} - \frac{1}{n+l.}-\frac{1}{n+l.-1}-\cdots-\frac{1}{n+l.-m+1}| \leq \epsilon$$

- $m, n, \min_j l_j (n \gg m)$ が大きければ上式は近似的に

$$|\frac{\min_j l_j}{\min_j l_j - m}| \approx |\frac{(\min_j l_j)(n+l.-m)}{(\min_j l_j - m)(n+l.)}| \leq \exp(\epsilon) \leftrightarrow (4)$$

考察）単純無作為復元抽出

- 復元抽出なら $n_j=0$ のセルのみ m_j の値域が n_j に依存。
- 模造機構として単純無作為復元抽出を採用した場合（多項分布）:

$$\mathsf{P}(\vec{m};\vec{n},\vec{\beta}) = \binom{m}{\vec{m}} \prod_{j=1}^{J} \left(\frac{n_j+\beta_j}{n+\beta.}\right)^{m_j}$$

  - $\beta_j > 0$ なら $n_j=0$ でも m_j の値域は変わらない。
- この時 ϵ-差分プライベート $\Leftrightarrow$

$$\left(1+\frac{1}{\min_j \beta_j}\right)^{m} \leq \exp(\epsilon) \quad (5)$$

- Machanavajjhala(2008) の例：$(m=$ 百万 $, \epsilon=7) \Rightarrow \beta_j \geq 142857$

単純無作為復元抽出の微分プライバシー

- 微分プライバシー $\Leftrightarrow \forall(\vec{m},\vec{n})$

$$|\frac{\partial \log \mathsf{P}(\vec{m};\vec{n})}{\partial n_j}| = |\frac{m_j}{n_j+\beta_j} - \frac{m}{n+\beta_.}| \leq \epsilon$$

- 上式左辺が最大化されるのは $n_j = 0, m_j = m$ の時なので、微分プライバシー $\Leftrightarrow$

$$|\frac{m}{\min_j \beta_j} - \frac{m}{n+\beta_.}| \leq \epsilon$$

- n が大きいときに上式は近似的に

$$\exp(\frac{m}{\min_j \beta_j}) \leq \exp(\epsilon)$$

- さらに $\min_j \beta_j$ が大きい時

$$(1+\frac{1}{\min_j \beta_j})^m \approx \exp(\frac{m}{\min_j \beta_j}) \leq \exp(\epsilon) \leftrightarrow (5)$$

考察）負の超幾何分布による模造

- Machanavajjhala (2008) の模造機構（負の超幾何分布）：

$$\mathsf{P}(\vec{m};\vec{n},\vec{\alpha}) = \binom{m}{\vec{m}} \frac{\Gamma(n+\alpha.)}{\Gamma(n+\alpha.+m)} \prod_{j=1}^{J} \frac{\Gamma(n_j+\alpha_j+m_j)}{\Gamma(n_j+\alpha_j)}$$

- この時 ϵ-差分プライベート $\Leftrightarrow$

$$\left|\frac{\min_j \alpha_j + m}{\min_j \alpha_j}\right| \leq \exp(\epsilon) \qquad (6)$$

- $\alpha_j = O(m)$ くらいで左辺を bound 出来る。
- Machanavajjhala(2008) の例：$(m = 百万\ , \epsilon = 7) \Rightarrow \alpha_j \geq 914$

負の超幾何分布の微分プライバシー

- 微分プライバシー $\Leftrightarrow \forall(\vec{m},\vec{n})$

$$\left|\frac{\partial \log \mathsf{P}(\vec{m};\vec{n})}{\partial n_j}\right| = \left|\frac{1}{n_j+\alpha_j}+\frac{1}{n_j+\alpha_j+1}+\cdots+\frac{1}{n_j+\alpha_j+m_j-1}\right.$$
$$\left.-\frac{1}{n+\alpha.}-\frac{1}{n+\alpha.+1}-\cdots-\frac{1}{n+\alpha.+m-1}\right| \le \epsilon$$

- 上式左辺が最大化されるのは $n_j=0, m_j=m$ の時なので、微分プライバシー $\Leftrightarrow$

$$\left|\frac{1}{\min_j \alpha_j}+\frac{1}{\min_j \alpha_j+1}+\cdots+\frac{1}{\min_j \alpha_j+m-1}\right.$$
$$\left.-\frac{1}{n+\alpha.}-\frac{1}{n+\alpha.+1}-\cdots-\frac{1}{n+\alpha.+m-1}\right| \le \epsilon$$

- $m, n, \min_j \alpha_j (n \gg m)$ が大きければ上式は近似的に

$$\left|\frac{\min_j \alpha_j + m}{\min_j \alpha_j}\right| \approx \left|\frac{(m+\min_j \alpha_j)(n+\alpha.)}{\min_j \alpha_j (n+m+\alpha.)}\right| \le \exp(\epsilon) \leftrightarrow (6)$$

考察）擬似多項分布による模造

- 擬似多項分布 (type 2) による模造機構：

$$\mathsf{P}(\vec{m};\vec{n},\vec{\gamma}) = \binom{m}{\vec{m}} \frac{1}{(n+\gamma.)(n+\gamma.+m)^{m-1}} \prod_{j=1}^{J} (n_j+\gamma_j)(n_j+\gamma_j+m_j)^{m_j-1}$$

- この時 ϵ-差分プライベート $\Leftrightarrow$

$$\left|\left(1+\frac{1}{\min_j \gamma_j}\right)\left(1+\frac{1}{\min_j \gamma_j+m}\right)^{m-1}\right| \le \exp(\epsilon) \qquad (7)$$

- この条件は m が大きいとき近似的に

$$(1+1/\min_j \gamma_j) \le \exp(\epsilon-1)$$

であり、γ_j を m 依存で膨らませる必要がない。

- Machanavajjhala(2008) の例：$(m=$ 百万 $, \epsilon=7) \Rightarrow \gamma_j \ge 0.00248491$

擬似多項分布の微分プライバシー

- 微分プライバシー $\Leftrightarrow \forall(\vec{m}, \vec{n})$

$$|\frac{\partial \log \mathsf{P}(\vec{m};\vec{n})}{\partial n_j}| = |\frac{m_j - 1}{n_j + \gamma_j + m_j} + \frac{1}{n_j + \gamma_j} - \frac{1}{n + \gamma_.} - \frac{m-1}{n + \gamma_. + m}| \leq \epsilon$$

- 上式左辺が最大化されるのは $n_j = 0, m_j = m$ の時なので、微分プライバシー $\Leftrightarrow$

$$|\frac{m-1}{\min_j \gamma_j + m} + \frac{1}{\min_j \gamma_j} - \frac{1}{n + \gamma_.} - \frac{m-1}{n + \gamma_. + m}| \leq \epsilon$$

- n が大きければ上式は近似的に

$$\exp(\frac{m-1}{\min_j \gamma_j + m} + \frac{1}{\min_j \gamma_j}) \leq \exp(\epsilon)$$

- さらに $\min_j \gamma_j$ が大きいとき左辺を近似して

$$(1 + \frac{1}{\min_j \gamma_j})(1 + \frac{1}{\min_j \gamma_j + m})^{m-1} \leftrightarrow (7)$$

微分プライバシーと差分プライバシー

- Machanavajjhala の枠組みだと、尤度の n に依存する部分が分母と分子でキャンセルして、DP 条件が n に依存しない。
  - 条件がシンプルになる。
  - n に依存しなくていいのか：母集団に占める一個体の重みが効かない？
- 微分プライバシーだと n が $n + \delta$ に変化する影響が n に依存して表れる。
- 以上の例では $n \to \infty$ として微分プライバシー条件から n を消去すれば差分プライバシー条件と近似的に等しい。

差分プライベートな模造機構の比較

- $(m = n = 100$ 万$, \epsilon = 7, J = 100$ 万$)$ かつ全てのセルで
  $l = 1000911.714, \alpha = 914, \beta = 142857, \gamma = 0.0024849$ が等しいとする。
- これらの条件の下で $n_j = 1$ 万のセル度数の期待値：

超幾何分布	多項分布	負の超幾何分布	擬似多項分布
1.00999	1.06999	11.9279	9975.22

  - $n_j = 0$ のセルに期待値を配るので $n_j > 0$ のセルの期待値は過小となる。
  - $n_j = 0$ のセルの保護をあきらめて n_j 依存で $(l_j, \alpha_j, \beta_j, \gamma_j)$ を決めれば、$n_j > 0$ のセルの期待値を不偏に出来る。
- 擬似多項分布はよさそうに見えるが、ほとんど実証していない。
  - Hu and Hoshino (2018, LNCS 11126)

擬似二項分布（青色）と二項分布（燈色）の比較

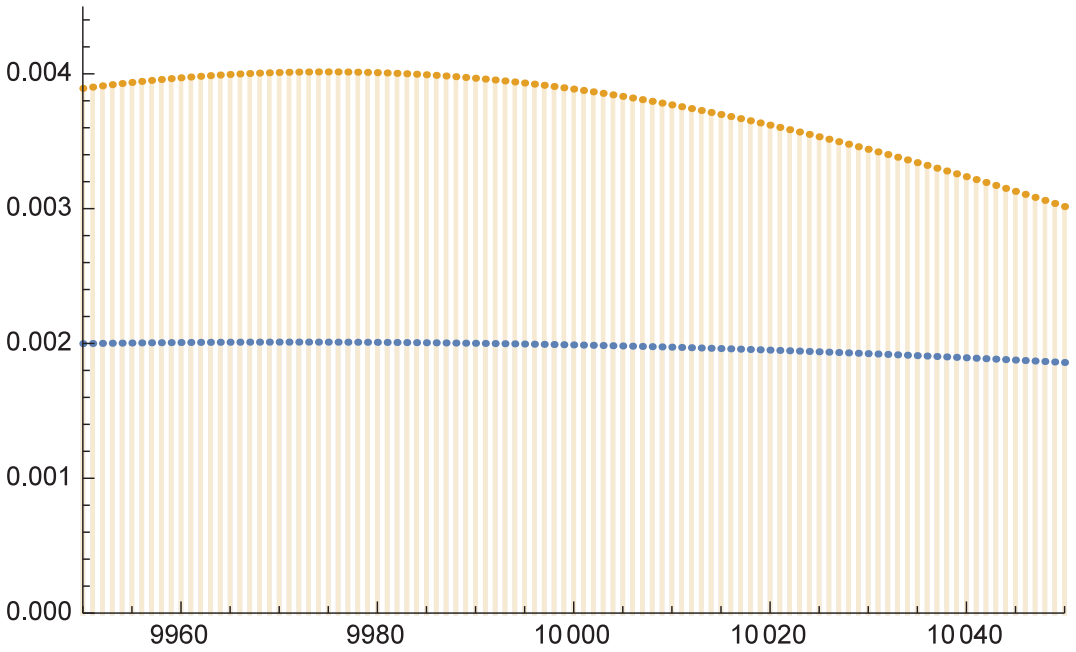


Figure 1: Probability Masses (Average 9975.22) of QB & Bin ($n = J = 1$ million, $\gamma = 0.0024849$)

擬似多項分布の紹介

The probability mass function (pmf) of QM$(\pi_1, \ldots, \pi_F; n, \beta)$:

$$p(y_1, \ldots, y_F) = \frac{n!}{y_1! \ldots y_F!} \frac{1}{(1+n\beta)^{n-1}} \prod_{f=1}^{F} \pi_f (\pi_f + y_f \beta)^{y_f - 1}$$

- y_f: the frequency of the f-th cell
- n: the given sum of F frequencies
- π_f: the f-th cell probability $(\pi_f \geq 0, \sum \pi_f = 1)$
- β: this parameter controls the variances of frequencies $(\beta \geq -\min_f \pi_f / n)$ *

- QM$(\pi_1, \ldots, \pi_F; n, 0) =$ Multinomial$(\pi_1, \ldots, \pi_F; n)$
- Our parameterization is different from the original (Consul and Mittal, 1977).

*Rejection sampling needs $\beta \geq 0$

Closure under the collapse of cells

The probability mass function of QM$(\pi_1, \ldots, \pi_F; n, \beta)$ is expressed as

$$p(y_1, \ldots, y_F) = \frac{\prod_{f=1}^{F} A_{y_f}(\pi_f, \beta)}{A_n(1, \beta)},$$

where

$$A_k(x, z) = x(x + kz)^{k-1}/k!, \quad k = 0, 1, 2 \ldots,$$

is called "Abel polynoials".

- Abel polynomials satisfy a convolution property (e.g., Charalambides, 2006, p.206):

$$A_n(x + y, z) = \sum_{k=0}^{n} A_k(x, z) A_{n-k}(y, z), \quad x \in \mathbb{R}, y \in \mathbb{R}, z \in \mathbb{R}.$$

$\Rightarrow$ QM is closed under the collapse of cells (H, 2009).

Conditional distribution method

Multivariate $\mathbf{Y}$ can be generated by the sequential sampling of univariate margins conditionally:

$$\mathbf{Y} \stackrel{d}{=} (Y_1|Y_2,\ldots,Y_F)\ldots(Y_{F-2}|Y_{F-1},Y_F)(Y_{F-1}|Y_F)Y_F$$

Theorem 1 *If* $\mathbf{Y} \sim QM(\pi_1,\ldots,\pi_F;n,\beta)$ *then* $(Y_f|Y_{f+1}=y_{f+1},\ldots,Y_F=y_F) \sim QB(\pi_f/(1-\sum_{g=f+1}^F \pi_g); n-\sum_{g=f+1}^F y_g, \beta)$ *for* $f=1,\ldots,F$.

- QB$(\pi;n,\beta)$ = QM$(\pi,1-\pi;n,\beta)$: "Quasi-Binomial (type 2)"
- Note that $\beta \geq -\min_{f\in[F]} \frac{\pi_f}{n} \geq$
  $-\min\left\{\frac{\pi_f}{(1-\sum_{g=f+1}^F \pi_g)(n-\sum_{g=f+1}^F y_g)}, \frac{\sum_{g=1}^{f-1}\pi_g}{(1-\sum_{g=f+1}^F \pi_g)(n-\sum_{g=f+1}^F y_g)}\right\}$
- Theorem holds even after exchanging the indices of cells.
  - Faster to sample when $\pi_1 \leq \pi_2 \leq \ldots \leq \pi_F$; see Ho et al. (1979)

- We can generate QM samples if we can generate QB samples.

Multi-stage sampling

Example (2-stage sampling) The first stage generates $\sum_{g=1}^f Y_g \sim$ QM$(\sum_{g=1}^f \pi_g; n, \beta)$. The second stage generates $(Y_1,\ldots,Y_f)$ given their sum. The second stage distribution is:

Theorem 2 *If* $\mathbf{Y} \sim QM(\pi_1,\ldots,\pi_F;n,\beta)$ *then for* $f=1,\ldots,F$ *and* $m=0,\ldots,n$,

$$(Y_1,\ldots,Y_f|\sum_{g=1}^f Y_g = m) \sim QM(\pi_1/(\sum_{g=1}^f \pi_g),\ldots,\pi_f/(\sum_{g=1}^f \pi_g); m, \beta).$$

- Note that $\beta \geq -\min_{f\in[F]} \frac{\pi_f}{n} \geq -\min_{h\in[f]} \frac{\pi_h/(\sum_{g=1}^f \pi_g)}{m}$

- A recursive argument validates multi-stage sampling from the QM distribution.
- Multi-stage sampling may be faster:
  - See Malefaki and Iliopoulos (2007)
  - It lowers the rejection rate of our sampling to be proposed.

Rejection sampling from QB

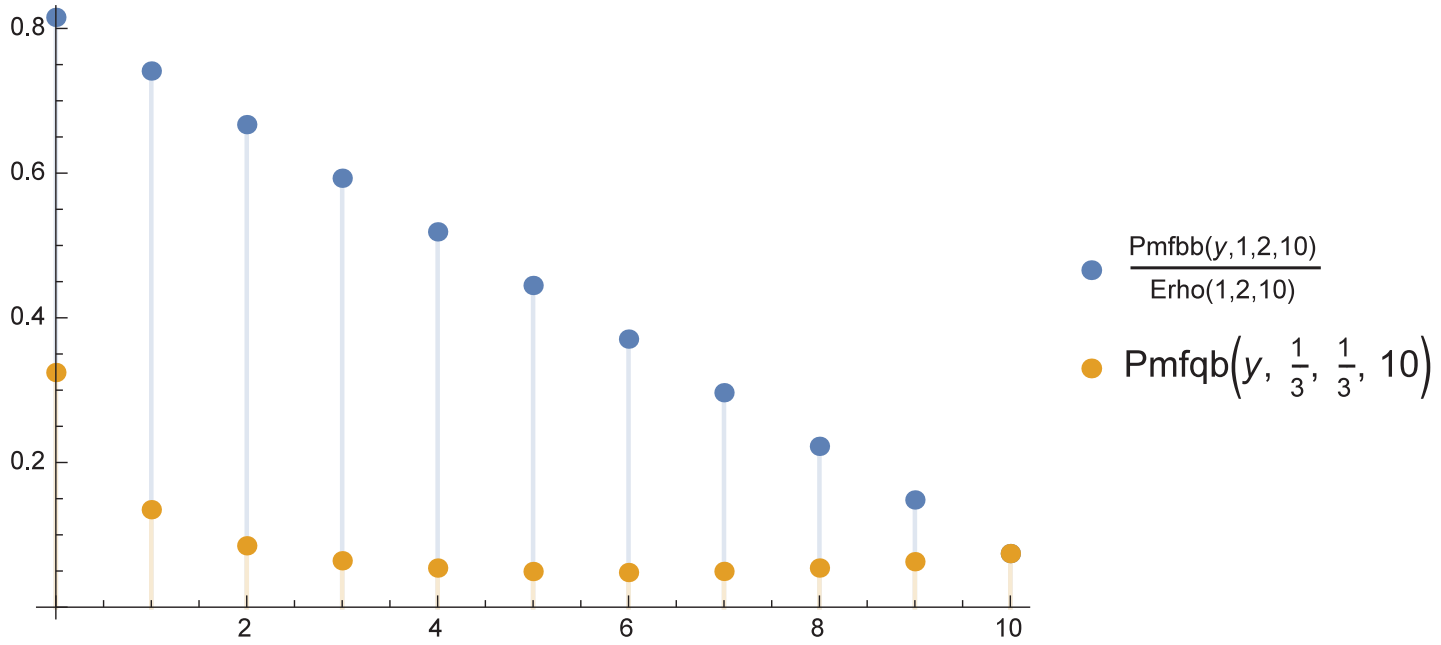


Figure 2: Dominating QB pmf by BB pmf (Average acceptance rate $\doteq 20\%$)

BB distribution

Beta(a_1, a_2) distribution is defined by the following density function:

$$f(p) = \frac{\Gamma(a_1)\Gamma(a_2)}{\Gamma(a_1+a_2)} p^{a_1-1}(1-p)^{a_2-1}, \quad 0 \le p \le 1,$$

where a_1, a_2 are usually required to be positive,

$Y \mid p \sim$ Binomial(n, π) then Y is called BB$(a_1, a_2; n)$ distributed, with pmf

$$p_{BB}(y) = \frac{n!}{y!(n-y)!} \frac{\Gamma(a.)}{\Gamma(a.+n)} \frac{\Gamma(a_1+y)}{\Gamma(a_1)} \frac{\Gamma(a_2+n-y)}{\Gamma(a_2)}, \quad y = 0, \ldots, n.$$

- Both BB and QB distributions belong to the class of CCP distributions (H, 2009).
- To equalize the "cell probabilities" of BB and QB, it has to be

$$\pi_1 = a_1/(a_1+a_2), \quad \beta = 1/(a_1+a_2)$$

Largest difference between pmfs

We need to find the minimum ratio of

$$\frac{p_{BB}(y)}{p_{QB}(y)} = \frac{\Gamma(a.+1)\Gamma(a_1+y)\Gamma(a_2+n-y)(a.+n)^{n-1}}{\Gamma(a.+n)\Gamma(a_1+1)\Gamma(a_2+1)(a_1+y)^{y-1}(a_2+(n-y))^{n-y-1}} \tag{8}$$

to decide the multiplication rate of p_{BB} to dominate p_{QB}.

Theorem 3 *Suppose that n is a positive integer, and a_1, a_2 are positive real numbers. Denote the value of y that minimizes Equation (8) by y^*. Then $y^* = n$ when $a_1 < a_2$, and $y^* = 0$ when $a_2 < a_1$. When $a_1 = a_2$, (8) is minimized at $y = 0$ and $y = n$.*

Efficiency of rejection sampling

By symmmetry, we only consider the case of $a_1 < a_2$. Acceptance rates at $y = 0, 1, \ldots, n$ are

$$\frac{\Gamma(a_1+n)\Gamma(a_2+1)}{\Gamma(a_1+y)\Gamma(a_2+n-y)} \frac{(a_1+y)^{y-1}(a_2+n-y)^{n-y-1}}{(a_1+n)^{n-1}} =: \rho(y).$$

Then

$$\mathsf{E}(\rho(Y)) = \min_y \frac{p_{BB}(y)}{p_{QB}(y)} = \frac{p_{BB}(n)}{p_{QB}(n)} = \frac{\Gamma(a.+1)\Gamma(a_1+n)}{\Gamma(a.+n)\Gamma(a_1+1)} \left(\frac{a.+n}{a_1+n}\right)^{n-1} =: r(a_1, a_2, n).$$

- $\lim_{a\to\infty} r(\pi a, (1-\pi)a, n) = 1$
- $\lim_{n\to\infty} r(a_1, a_2, n) = O(n^{-a_2})$

Table 1: Average Acceptance Rates: $r(0.1/\beta, 0.9/\beta, n)$

n	β									
	2^{-1}	2^{-2}	2^{-3}	2^{-4}	2^{-5}	2^{-6}	2^{-7}	2^{-8}	2^{-9}	2^{-10}
10^1	0.13	0.06	0.03	0.02	0.03	0.06	0.14	0.29	0.50	0.69
10^2	0.00	0.00	0.01	0.07	0.26	0.51	0.71	0.84	0.92	0.96
10^3	0.12	0.34	0.59	0.77	0.87	0.94	0.97	0.98	0.99	1.00
10^4	0.81	0.90	0.95	0.97	0.99	0.99	1.00	1.00	1.00	1.00
10^5	0.98	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
10^6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
10^7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
10^8	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84
10^9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Algorithm 1 (Rejection sampling from the QB distribution) *The following procedure generates a sample from QB*$(a_1/(a_1+a_2); n, 1/(a_1+a_2))$ *for a positive integer* n.

When $0 < a_1 < a_2$,

1. *Generate* $p \sim$ *Beta*(a_1, a_2)
2. *Generate* $y|p \sim$ *Binomial*(n, p)
3. *Generate* $u \sim$ *U*$(0, 1)$
4. *If* $u > \rho(y)$ *then goto 1*
5. *Output* y.

When $0 < a_2 < a_1$,

1. *Swap* a_2 *and* a_1.
2. *Generate* $p \sim$ *Beta*(a_1, a_2)
3. *Generate* $y|p \sim$ *Binomial*(n, p)
4. *Generate* $u \sim$ *U*$(0, 1)$
5. *If* $u > \rho(y)$ *then goto 2*
6. *Output* $n - y$.

統計的推測精度の管理に向けて

- 差分プライバシーというきつい基準の下でも、擬似多項分布は有用なデータを生成できるかもしれない。
- 微分プライバシー概念の方が含意が明瞭。昨年お話したようにベル多項式を用いると見通しが良くなる。
- 差分プライバシーは全ての母数の下での安全性を考えるが、所与の母数（発行したいデータ）の下での安全性でよいはず。
  - $\vec{n}$ について最大化しない条件付き差分/微分プライバシーを押さえればよい。
  - 条件付きで議論すれば、不偏性のような性質が導入可能。$\vec{n}$ を非確率的に操作するマスクの評価に適する。

多次元非定常時系列への新しいフィルタリング法[1]

2019 年 1 月

国友直人 (明治大学)

1 はじめに

- 多くのマクロ経済データにおいては非定常性、季節性、定常的な循環、観測ノイズなどがしばしば観察される。例えば GDP 統計の季節調整・設備投資系列の予測 (内閣府)、複数のマクロ消費系列からの合成指標の構成 (統計局) など、幾つかの具体的問題がある。

- 政府統計で広く利用されている季節調整法の X-12-ARIMA では一次元 Box-Jenkins 法 (Reg-ARIMA モデル) が古典的な移動平均法 (Moving Average) とともに (十分な理解があるといえない中でも) 広く利用されている。

- 統計数理研究所で開発された赤池・石黒の BAYSEA、北川の DECOMP（Kitagawa(2010)）においてはガウス尤度関数に基づくカルマン・フィルタリングにより 1 次元時系列の成分分解が実用化されている。多次元化はかなり困難である。

- 多くのマクロ時系列では全く同一ではないが似たような季節性や (景気) 循環変動などの共変動が観察されている。Engle-Granger(1987) では非定常系列において VAR(vector AR) による共和分 (co-integration) 分析を提唱しているが、(現実的な) 季節性や観測ノイズの存在は無視している。

- 四半期データ (20 年) では 80、月次データでは 120 程度の時系列データを想定、時系列としては小標本データである。

- 他方、Kunitomo, Sato and Kurisu (2018) では金融高頻度データ分析において分離最尤推定 (SIML) 法を提案している。（*Separating Information Maximum Likelihood Estimation for High Frequency Financial Data*, Springer）本稿で説明するのは SIML をマクロ (離散時間) 時系列に応用する方法であり、Nishimura, Sato and Takahashi (2019) による smoothing-method の一般化に対応している。

- 本稿では多次元非定常時系列の分布にあまり依存しないフィルタリング法を考察したので報告する。応用可能性は広範におよぶが、マクロ・トレンド抽出や季節調整の問題などが実例である。

[1] 2019-2-5. 佐藤整尚氏との共同研究 Kunitomo-Sato(2017), Kunitomo-Sato(2019) に基づく。高橋明彦氏からの指摘に感謝する。

2　マクロ経済データの例

マクロ経済データとして重要な系列である GDP 統計における四半期の実質 GDP と実質消費の二次元時系列、マクロ消費系列の月次系列 (家計調査、消動、第三次指数) を例として議論する。

3　非定常変数誤差モデル

簡単な離散時間の変数誤差モデル

$$\mathbf{y}_i = \mathbf{x}_i + \mathbf{v}_i \ \ (i = 1, \cdots, n), \tag{3.1}$$

を考察する。ここで $\mathbf{x}_i$ $(i = 1, \cdots, n)$ は非定常 I(1) 過程であり

$$\Delta \mathbf{x}_i = (1 - \mathcal{L})\mathbf{x}_i = \mathbf{v}_i^{(x)} \tag{3.2}$$

ラグ作用素 $\mathcal{L}\mathbf{x}_i = \mathbf{x}_{i-1}$, $\Delta = 1 - \mathcal{L}$,

$$\mathbf{v}_i = \sum_{j=0}^{\infty} \mathbf{C}_j^{(v)} \mathbf{e}_{i-j}^{(v)} \ , \tag{3.3}$$

$$\mathbf{v}_i^{(x)} = \sum_{j=0}^{\infty} \mathbf{C}_j^{(x)} \mathbf{e}_{i-j}^{(x)} \ , \tag{3.4}$$

$\mathbf{e}_i^{(v)}$, $\mathbf{e}_i^{(x)}$, は i.i.d. 系列、$\mathbf{E}(\mathbf{e}_i^{(v)}) = \mathbf{0}$, $\mathbf{E}(\mathbf{e}_i^{(x)}) = \mathbf{0}$, $\mathbf{E}(\mathbf{e}_i^{(v)}\mathbf{e}_i^{(v)\prime}) = \mathbf{\Sigma}_e^{(v)}$ (正定). $\mathbf{E}(\mathbf{e}_i^{(x)}\mathbf{e}_i^{(x)\prime}) = \mathbf{\Sigma}_e^{(x)}$ (非負定), 係数和の絶対値収束性を仮定する。

さらに $\mathbf{f}_{\Delta x}(\mu)$, $\mathbf{f}_v(\mu)$ をそれぞれ $\Delta\mathbf{x}_i$,$\mathbf{v}_i$ $(i = 1, \cdots, n)$ の $(p \times p)$ スペクトル密度行列

$$\mathbf{f}_{\Delta x}(\mu) = \frac{1}{\pi}(\sum_{j=0}^{\infty} \mathbf{C}_j^{(x)} e^{2\pi i \mu j}) \mathbf{\Sigma}_e^{(x)} (\sum_{j=0}^{\infty} \mathbf{C}_j^{(x)\prime} e^{-2\pi i \mu j}) \quad (-\frac{1}{2} \leq \mu \leq \frac{1}{2}) \ , \tag{3.5}$$

$$\mathbf{f}_v(\mu) = \frac{1}{\pi}(\sum_{j=0}^{\infty} \mathbf{C}_j^{(v)} e^{2\pi i \mu j}) \mathbf{\Sigma}_e^{(v)} (\sum_{j=0}^{\infty} \mathbf{C}_j^{(v)\prime} e^{-2\pi i \mu j}) \quad (-\frac{1}{2} \leq \mu \leq \frac{1}{2}) \tag{3.6}$$

基準化 $\mathbf{C}_0^{(x)} = \mathbf{C}_0^{(s)} = \mathbf{C}_0^{(v)} = \mathbf{I}_p$, $i^2 = -1$。階差変換 $\Delta\mathbf{y}_i$ $(= \mathbf{y}_i - \mathbf{y}_{i-1})$ の $p \times p$ スペクトル密度行列は

$$\mathbf{f}_{\Delta y}(\mu) = \mathbf{f}_{\Delta x}(\mu) + (1 - e^{2\pi i \mu}) f_v(\mu) (1 - e^{-2\pi i \mu}) \tag{3.7}$$

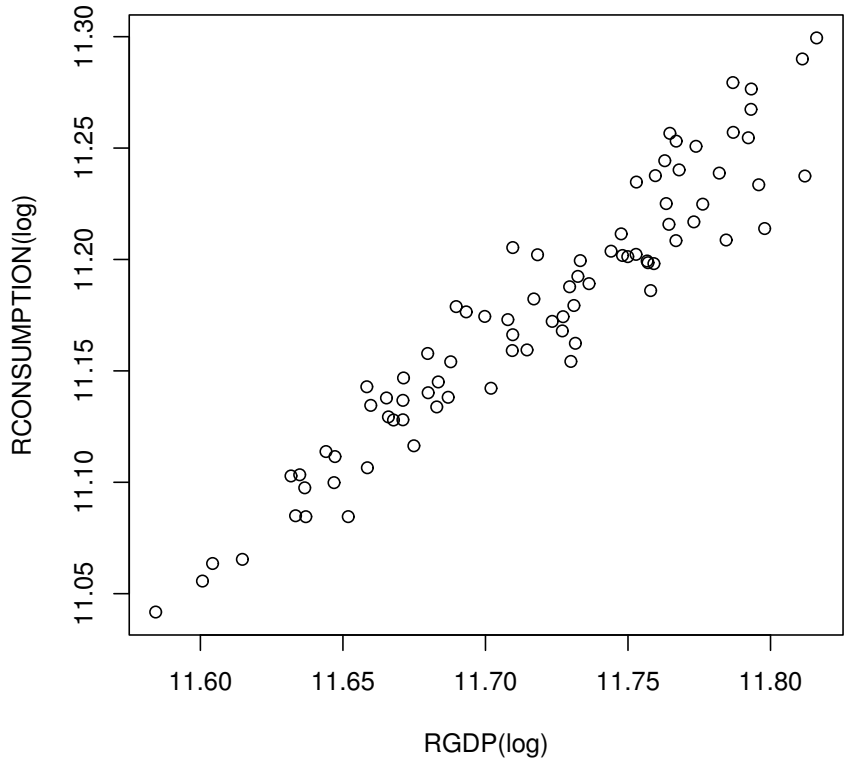


図 2.1 : 実質 GDP vs. 実質消費

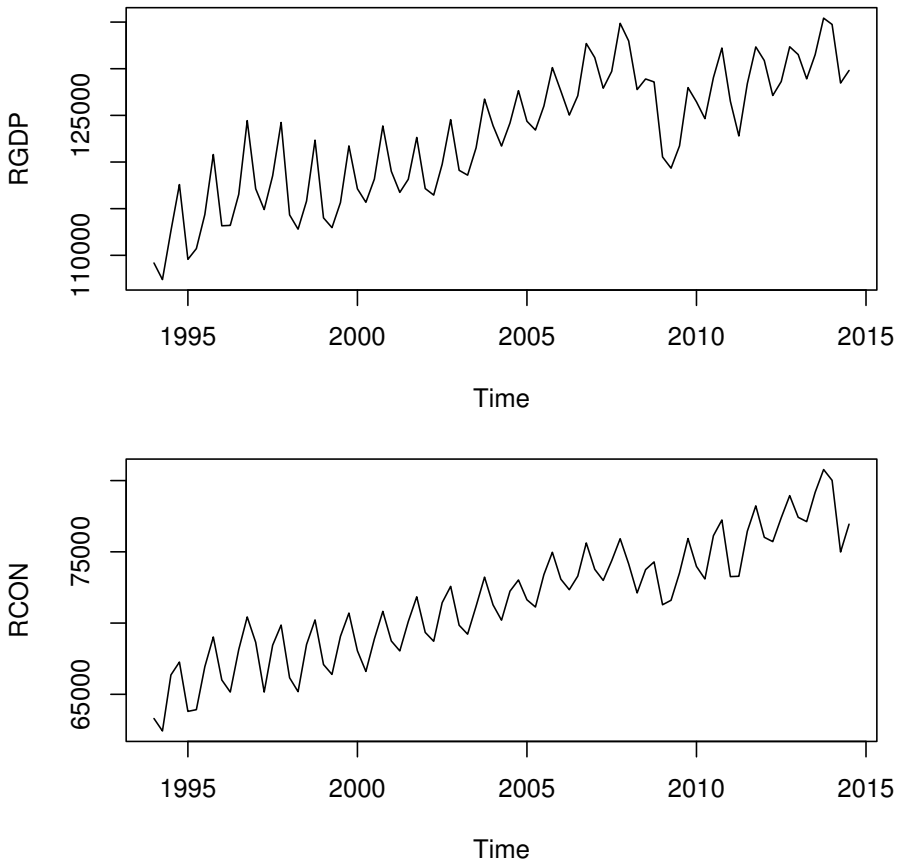


図 2.2 : 実質 GDP vs. 実質消費

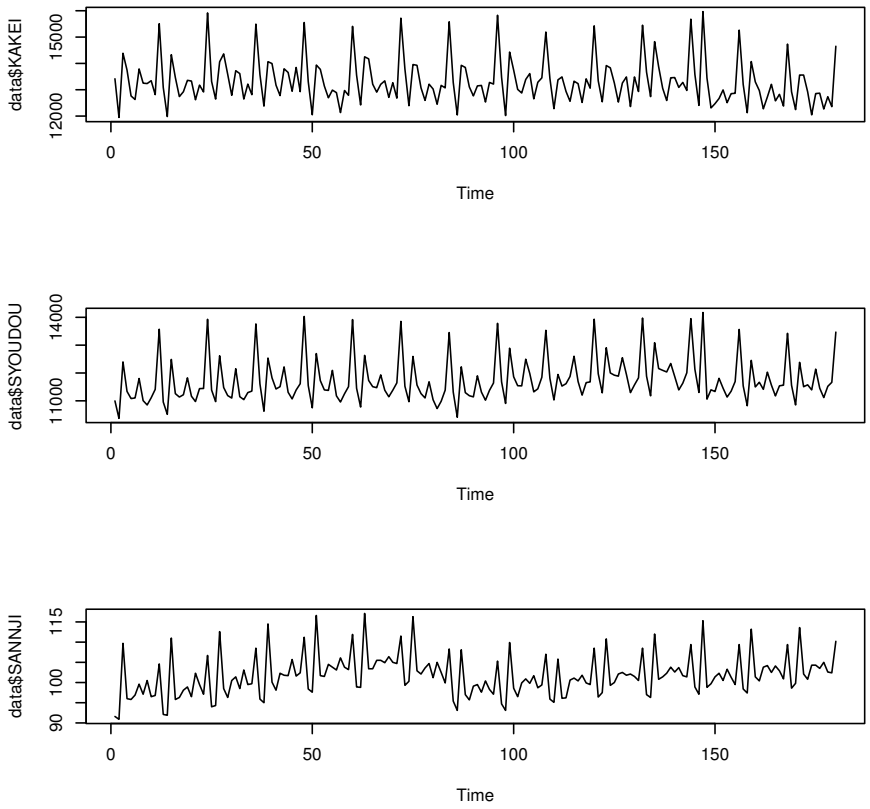

data$KAKEI
12000
15000
0
50
100
150
Time
data$SYOUDOU
11000
14000
0
50
100
150
Time
data$SANNJI
90
100
115
0
50
100
150
Time

図 3.1： マクロ消費系列

と表現される。各成分の長期分散共分散行列 $(g, h = 1, \cdots, p)$ は

$$\mathbf{\Omega}_x = \mathbf{f}_{\Delta x}(0) \ (= (\omega_{gh}^{(x)})) \ , \tag{3.8}$$

$$\mathbf{\Omega}_v = f_v(0) \ = (\omega_{gh}^{(v)}) \tag{3.9}$$

で与えられる。

季節性を含む加法モデル

$$\mathbf{y}_i = \mathbf{x}_i + \mathbf{s}_i + \mathbf{v}_i \ \ (i = 1, \cdots, n), \tag{3.10}$$

において正整数 s $(s > 1)$, N, $n = sN$ とした季節要素を $\mathbf{s}_i$ $(i = 1, \cdots, n)$ 非定常確率過程 $\Delta \mathbf{s}_i = (1 - \mathcal{L})\mathbf{s}_i = \mathbf{v}_i^{(s)}$,

$$\mathbf{v}_i^{(s)} = \sum_{j=0}^{\infty} \mathbf{C}_{js}^{(s)} \mathbf{e}_{i-sj}^{(s)} \ , \tag{3.11}$$

$\mathbf{e}_i^{(s)}$ は i.i.d. 系列、$\mathbf{E}(\mathbf{e}_i^{(s)}) = \mathbf{0}$, $\mathbf{E}(\mathbf{e}_i^{(s)} \mathbf{e}_i^{(s)'}) = \mathbf{\Sigma}_e^{(s)}$、係数和の絶対値収束性を仮定する。

さらに $\mathbf{f}_{\Delta s}(\mu)$ を $\Delta \mathbf{s}_i$ の $(p \times p)$ スペクトル密度行列とすると

$$\mathbf{f}_{\Delta s}(\mu) = \frac{1}{\pi} (\sum_{j=0}^{\infty} \mathbf{C}_{sj}^{(s)} e^{2\pi i \mu s j}) \mathbf{\Sigma}_e^{(s)} (\sum_{j=0}^{\infty} \mathbf{C}_{sj}^{(s)'} e^{-2\pi i \mu s j}) \quad (-\frac{1}{2} \leq \mu \leq \frac{1}{2}) \tag{3.12}$$

により与えられる。ここで基準化 $\mathbf{C}_0^{(s)} = \mathbf{I}_p$, $i^2 = -1$。階差変換 $\Delta \mathbf{y}_i$ $(= \mathbf{y}_i - \mathbf{y}_{i-1})$ の $p \times p$ スペクトル密度行列は

$$\mathbf{f}_{\Delta y}(\mu) = \mathbf{f}_{\Delta x}(\mu) + \mathbf{f}_{\Delta s}(\mu) + (1 - e^{2\pi i \mu}) f_v(\mu) (1 - e^{-2\pi i \mu}) \ . \tag{3.13}$$

長期分散共分散行列 $(g, h = 1, \cdots, p)$ は

$$\mathbf{\Omega}_s = \mathbf{f}_{\Delta s}(\frac{1}{s}) \ (= (\omega_{gh}^{(s)})) \tag{3.14}$$

で与えられる。

4 $\mathbf{K}_n$ 変換と $\mathbf{Z}_n$ 過程

データ行列に対して $\mathbf{K}_n-$ 変換（$\mathbf{Y}_n$ より $\mathbf{Z}_n$ $(= (\mathbf{z}_k'))$）は

$$\mathbf{Z}_n = \mathbf{K}_n \left(\mathbf{Y}_n - \bar{\mathbf{Y}}_0 \right) \ , \mathbf{K}_n = \mathbf{P}_n \mathbf{C}_n^{-1} \ , \tag{4.15}$$

により定義。ただし

$$\mathbf{C}_n^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & 0 & \cdots \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}_{n\times n} , \tag{4.16}$$

$$\mathbf{P}_n = (p_{jk}^{(n)}) \ , \ p_{jk}^{(n)} = \sqrt{\frac{2}{n+\frac{1}{2}}} \cos\left[\frac{2\pi}{2n+1}(k-\frac{1}{2})(j-\frac{1}{2})\right] \ . \tag{4.17}$$

このときスペクトル分解により $\mathbf{C}_n^{-1}\mathbf{C}_n^{\prime -1} = \mathbf{P}_n\mathbf{D}_n\mathbf{P}_n^{'}$ $\mathbf{D}_n$ は対角行列、第 (k,k) 要素は $d_k = 2[1-\cos(\pi(\frac{2k-1}{2n+1}))] = 4\sin^2(\frac{\pi}{2}(\frac{2k-1}{2n+1}))$ $(k = 1, \cdots, n)$。
さらに $\mathbf{K}_n-$ 変換された系列 $\mathbf{Z}_n$ に対するフィルタリング (あるいは Nishimura, et al. (2019) で提案された smoothing) 法を考える。$m \times n$ 選択行列 $\mathbf{J}_m = (\mathbf{I}_m, \mathbf{O})$ として $n \times p$ 行列

$$\hat{\mathbf{X}}_n = \mathbf{C}_n\mathbf{P}_n^{'}\mathbf{J}_m^{'}\mathbf{J}_m\mathbf{P}_n\mathbf{C}_n^{-1}(\mathbf{Y}_n - \bar{\mathbf{Y}}_0) \tag{4.18}$$

ただし

$$\mathbf{Z}_n = \mathbf{P}_n\mathbf{C}_n^{-1}(\mathbf{Y}_n - \bar{\mathbf{Y}}_0) \tag{4.19}$$

$n \times n$ 変換

$$\mathbf{Q}_n = \mathbf{P}_n\mathbf{J}_m^{'}\mathbf{J}_m\mathbf{P}_n \tag{4.20}$$

で与えられる。

次により一般の $\mathbf{K}_n-$ 変換された系列 $\mathbf{Z}_n$ に対するフィルタリングを考える。例えば季節性は離散時間の季節周期 s (> 1) と理解できる。$m_2 \times [m_1+m_2+(n-m_1-m_2)]$ 選択行列 $\mathbf{J}_{m_1,m_2,n} = (\mathbf{O}, \mathbf{I}_{m_2}, \mathbf{O})$, $n \times p$ 行列

$$\hat{\mathbf{W}}_n = \mathbf{C}_n\mathbf{P}_n^{'}\mathbf{J}_{m_1,m_2,n}^{'}\mathbf{J}_{m_1,m_2,n}\mathbf{P}_n\mathbf{C}_n^{-1}(\mathbf{Y}_n - \bar{\mathbf{Y}}_0) \tag{4.21}$$

$n \times n$ 行列

$$\mathbf{Q}_n = \mathbf{P}_n\mathbf{J}_{m_1,m_2,n}^{'}\mathbf{J}_{m_1,m_2,n}\mathbf{P}_n \ . \tag{4.22}$$

を利用する。このとき $(j, j^{'})$-要素 $\mathbf{Q}_n = \mathbf{P}_n\mathbf{J}_{m_1,m_2,n}^{'}\mathbf{J}_{m_1,m_2,n}\mathbf{P}_n$ $(= (q_{j,j'}))$ は

$$\begin{aligned} q_{j,j} &= \frac{2m_2}{2n+1} + \frac{1}{2n+1}\left[\frac{\sin\frac{2(m_1+m_2)\pi}{2n+1}(2j-1) - \sin\frac{2(m_1)\pi}{2n+1}(2j-1)}{\sin\frac{\pi}{2n+1}(2j-1)}\right] , \\ q_{i,j'} &= \frac{1}{2n+1}\left[\frac{\sin\frac{2(m_1 1+m_2)\pi}{2n+1}(j+j^{'}-1) - \sin\frac{2(m_1)\pi}{2n+1}(j+j^{'}-1)}{\sin\frac{\pi}{2n+1}(j+j^{'}-1)}\right. \\ &\qquad \left. + \frac{\sin\frac{2(m_1+m_2)\pi}{2n+1}(j-j^{'}) - \sin\frac{2(m_1)\pi}{2n+1}(j-j^{'})}{\sin\frac{\pi}{2n+1}(j-j^{'})}\right] \quad (j \neq j^{'}) \end{aligned}$$

となる。

5　データの直交分解

記号 $\theta_{jk} = \frac{2\pi}{2n+1}(j-\frac{1}{2})(k-\frac{1}{2})$,

$$p_{jk}^{(n)} = \frac{1}{\sqrt{2n+1}}(e^{i\theta_{jk}} + e^{-i\theta_{jk}}) \tag{5.23}$$

を利用して

$$\Delta_\lambda \mathbf{z}^{(n)}(\lambda_k^{(n)}) = \sum_{j=1}^{n} p_{jk}^{(n)} \mathbf{r}_j^{(n)} \ , \ \mathbf{r}_j^{(n)} = \mathbf{y}_j - \mathbf{y}_{j-1} \ , \tag{5.24}$$

と表現すると、変換系列 $\mathbf{Z}_n$ はデータのある種の実フーリエ変換 $\Delta_\lambda \mathbf{z}^{(n)}(\lambda_k^{(n)})$ $(k = 1, \cdots, n)$ はデータの周波数 $\lambda_k^{(n)}$ $(= (k-1/2)/(2n+1))$, におけるフーリエ変換、データ直交増分過程である。

定理 5.1 : (離散時間) 確率過程 $\mathbf{r}_j$ $(j = 1, \cdots, n)$ はエルゴード的定常過程、$\mathbf{\Gamma}(h) = \mathcal{E}(\mathbf{r}_j \mathbf{r}_{j-h}^{'})$ は

$$\sum_{h=0}^{\infty} \|\mathbf{\Gamma}(h)\| < \infty \ . \tag{5.25}$$

を満たす (有界性) ことを仮定する。
(i) $\Delta_\lambda \mathbf{z}^{(n)}(\lambda_k^{(n)}) = \sum_{j=1}^{n} p_{jk}^{(n)} \mathbf{r}_j^{(n)}$, $\mathbf{r}_j^{(n)}$ がエルゴード的定常過程で $\mathcal{E}[\mathbf{r}_j] = \mathbf{0}$, 対称化実スペクトル密度行列

$$\mathbf{f}_{SR}(\lambda) = \mathbf{\Gamma}(0) + \sum_{h=1}^{\infty} \cos(2\pi h\lambda)[\mathbf{\Gamma}(h) + \mathbf{\Gamma}(-h)] \ , \tag{5.26}$$

は正定符号、有界性を仮定。$\lambda_k^{(n)} \to s$, $\lambda_{k'}^{(n)} \to t (0 < s < t < \frac{1}{2})$. $n \longrightarrow \infty$ のとき

$$\begin{bmatrix} \Delta_\lambda \mathbf{z}^{(n)}(\lambda_k^{(n)}) \\ \Delta_\lambda \mathbf{z}^{(n)}(\lambda_{k'}^{(n)}) \end{bmatrix} \xrightarrow{w} N_{2p}\left[\mathbf{0}, \begin{bmatrix} \mathbf{f}_{SR}(s) & \mathbf{0} \\ \mathbf{0} & \mathbf{f}_{SR}(t) \end{bmatrix}\right] . \tag{5.27}$$

(ii) 増分過程 $\mathbf{Z}_n(t) - \mathbf{Z}_n(s) = \frac{1}{\sqrt{n}} \sum_{k=[sn]}^{[tn]} \sum_{j=1}^{n} p_{jk}^{(n)} \mathbf{r}_j^{(n)}$ $(0 < s < t < 1)$. $n \longrightarrow \infty$ のとき

$$\mathbf{Z}_n(t) - \mathbf{Z}_n(s) \xrightarrow{w} N_p\left[\mathbf{0}, F_{SR}(t) - F_{SR}(s)\right] \ , \tag{5.28}$$

ただし $F_{SR}(t)$ は $p \times p$ (対称化実) スペクトル分布行列

$$F_{SR}(t) = \int_0^t f_{SR}(\lambda) d\lambda \ . \tag{5.29}$$

で与えられる。

(実) スペクトル経験分布 (行列) は

$$\mathbf{F}_{SR,n}(t) = \frac{1}{n}\sum_{k\leq[2nt]}(\Delta_t \mathbf{z}_k^{(n)}(\lambda_k^{(n)}))(\Delta_t \mathbf{z}_k^{(n)}(\lambda_k^{(n)}))' \ . \tag{5.30}$$

となる。

KS(2017) における SIML 推定量は

$$\mathbf{G}_m = \frac{1}{m}\sum_{k=1}^{m}\Delta_0 \mathbf{z}_k^{(n)}(\lambda_k)\Delta_0 \mathbf{z}_k^{(n)}(\lambda_k)' \ . \tag{5.31}$$

より一般には

$$\mathbf{G}_m(t) = \frac{1}{m}\sum_{k=[2nt]-\frac{m_n}{2}+1}^{[2nt]+\frac{m_n}{2}}(\Delta_t \mathbf{z}_k^{(n)}(\lambda_k^{(n)}))(\Delta_t \mathbf{z}_k^{(n)}(\lambda_k^{(n)}))' \ . \tag{5.32}$$

定理 5.2 : (離散時間) 確率過程 $\mathbf{r}_j$ $(j = 1, \cdots, n)$ がエルゴード的定常過程、$\mathbf{\Gamma}(h) = \mathcal{E}(\mathbf{r}_j \mathbf{r}_{j-h}^{'})$ とする（有界性を仮定)。
(i) $m_n = [n^\alpha]$ $(0 < \alpha < 1)$ とおくと、任意の $t \in (0, \frac{1}{2})$ に対し $n \longrightarrow \infty$ のとき

$$\mathbf{G}_m(t) \xrightarrow{p} \mathbf{f}_{SR}(t) \tag{5.33}$$

(ii)$n \longrightarrow \infty$ のとき

$$\mathbf{F}_{SR,m}(t) \xrightarrow{p} \mathbf{F}_{SR}(t) \ . \tag{5.34}$$

6 Filtering と Smoothing 法

伝統的時系列論ではスペクトル分布 F を持つ定常 (離散時間) ベクトル過程 $\mathbf{r}_k^*$ に対して右連続な直交増分 (ベクトル, 複素数値) 過程 $\mathbf{z}^*(\lambda)$ $(-1/2 \leq \lambda \leq 1/2)$ が存在し、

$$\mathbf{r}_k^* = \int_{(-1/2,1/2]} e^{i2\pi ik\nu} d\mathbf{z}^*(\nu) \ (k = 1, \cdots, n). \tag{6.35}$$

と表現されることが知られている。(Doob (1953), Hannan (1971), Brockwell and Davis (1990) などを参照.)

本稿におけるような実ベクトル値をとる確率過程 (理論値) は

$$\mathbf{r}_k^{(u)} = \int_{(0,1/2]} \cos(2\pi k\nu) h_u(\cos(2\pi k\nu) d\mathbf{z}(\nu) \ (k = 1, \cdots, n) \tag{6.36}$$

$(u = x\ , u = s)$、データからの直交過程による (離散時間) 平滑化 (smoothing) 値は

$$\mathbf{r}_k^{(n,u)} = \int_{(0,1/2]} \cos(2\pi k\nu) h_{u,n}(\cos(2\pi k\nu) d\mathbf{z}_n(\nu)\ (k = 1, \cdots, n) \tag{6.37}$$

$(u = x\ , u = s)$ と表現される（$h(\cdot)$ はカーネル関数である)。

7 実例と議論

幾つかのマクロ系列を利用して本稿で考察しているフィルタリング法の妥当性が高いことが分かった。伝統的な時系列解析では定常過程にもとづくスペクトル密度の推定に高い関心を示していたが、その方法では非定常系列では上手くいかない理由が明らかになる。また例えばBox-Jenkin 法で推奨されている季節階差法には基本的な問題がある、ことなどがここでの分析から分かる。ここでは幾つかのデータ分析の例示にとどめる。

8 まとめと展望

- 多くのマクロ経済データにおいては非定常性、季節性、定常的な循環、観測ノイズなどが。例えばGDP 統計の季節調整・系列の合成 (内閣府)、複数のマクロ消費系列から総合指標の作成 (統計局) など、幾つかの問題がある。
- 季節調整法の X-12-ARIMA、DECOMP、VAR による共和分などはいずれも改善する必要がある。
- 本稿では多次元非定常時系列の分布にあまり依存しないマクロ SIML 法にもとづく新しいフィルタリング法を議論した。多次元非定常系列の実務的課題、例えば、トレンド抽出、季節調整、集計系列の合成、などへの応用可能性は広範と考えられる。
- 多期間の予測誤差を最小化するような最適なカーネル $h_{u,n}(\cdot)$ の選択を考察中、また離散時間確率過程 $\Delta_\lambda \mathbf{z}^{(n)}(\lambda_k^{(n)})$ は基準化すると右連続な直交増分 (ベクトル, 実数値) 過程 $\mathbf{z}(\lambda)$ $(0 \leq \lambda \leq 1/2)$ に弱収束するが、filtering、smoothing の理論的誤差評価を考察中である。

References

[1] Anderson, T.W. (1971), *The Statistical Analysis of Time Series*, John-Wiley.

[2] Anderson, T.W. (1984), "Estimating Linear Statistical Relationships," *Annals of Statistics*, 12, 1-45.

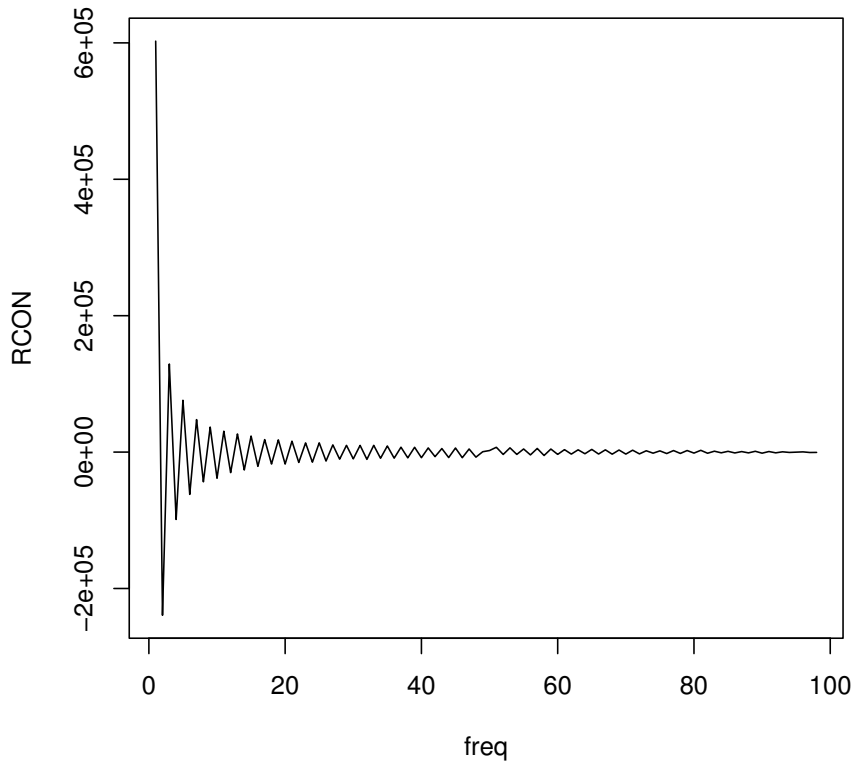


図 7.1： 実質消費 (原系列) の z_n 過程

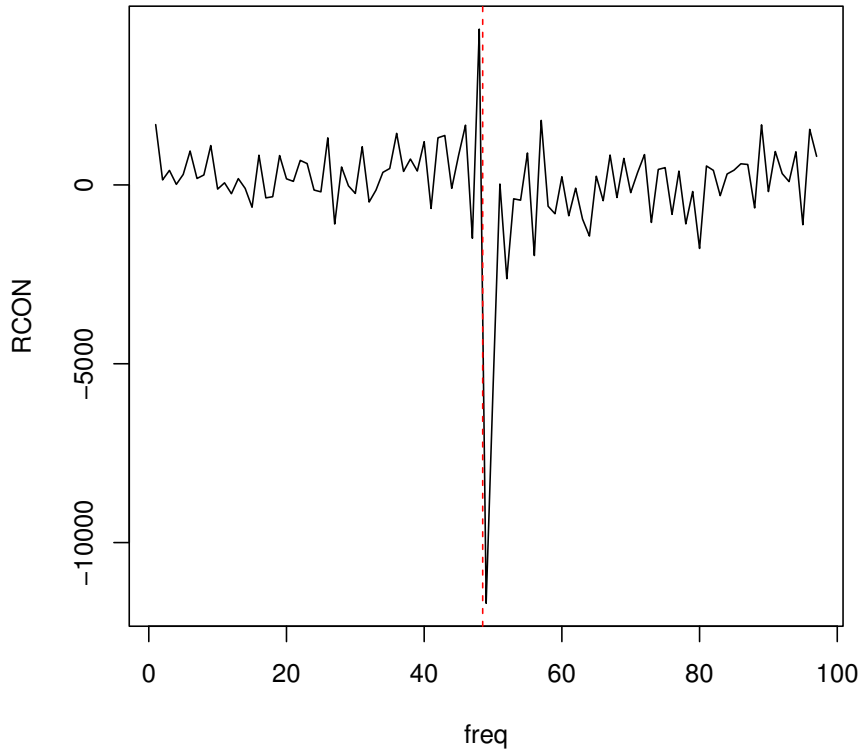


図 7.2 : 実質消費 (階差) の z_n 過程

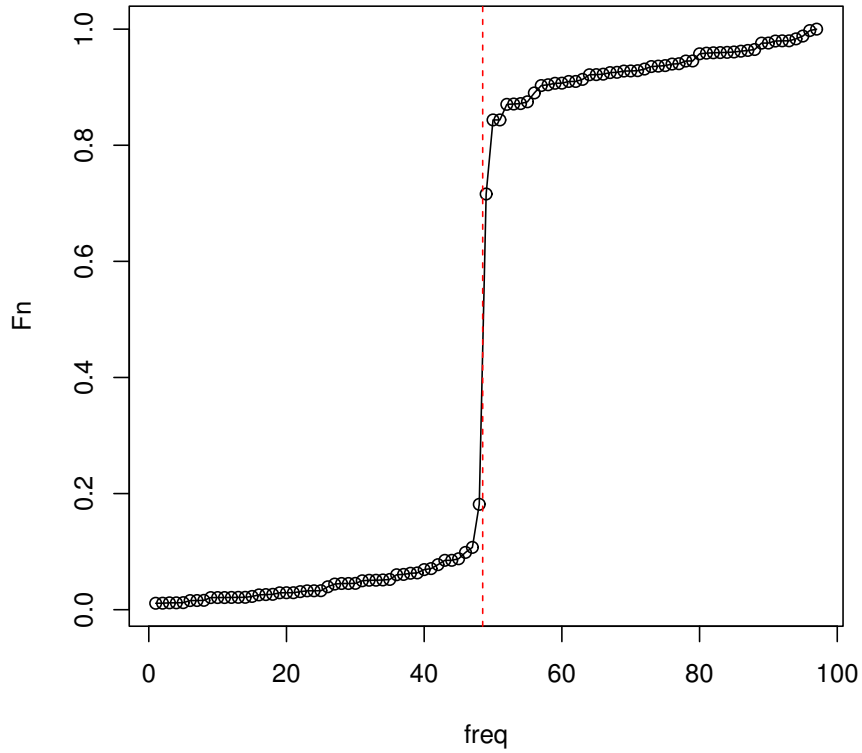


図 7.3 : 実質消費 (経験スペクトル分布)

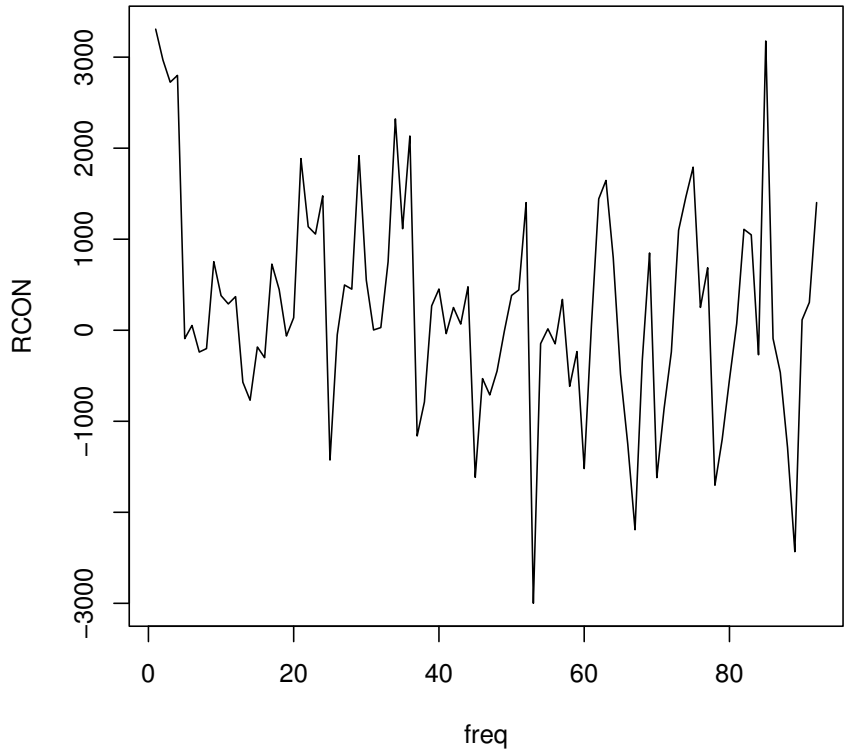


図 7.4 : 実質消費 (季節階差) の z_n 過程

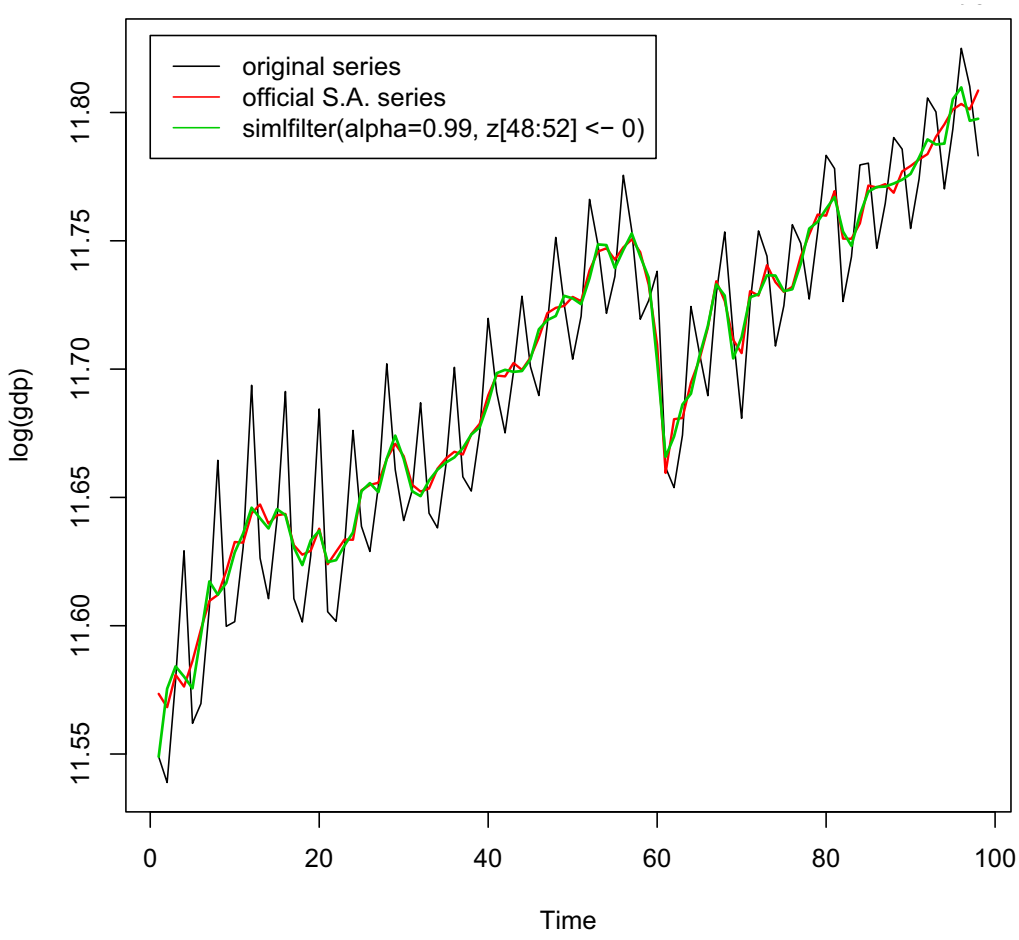

original series
official S.A. series
simlfilter(alpha=0.99, z[48:52] <- 0)
log(gdp)
11.55
11.60
11.65
11.70
11.75
11.80
0
20
40
60
80
100
Time

図 7.5：実質 GDP

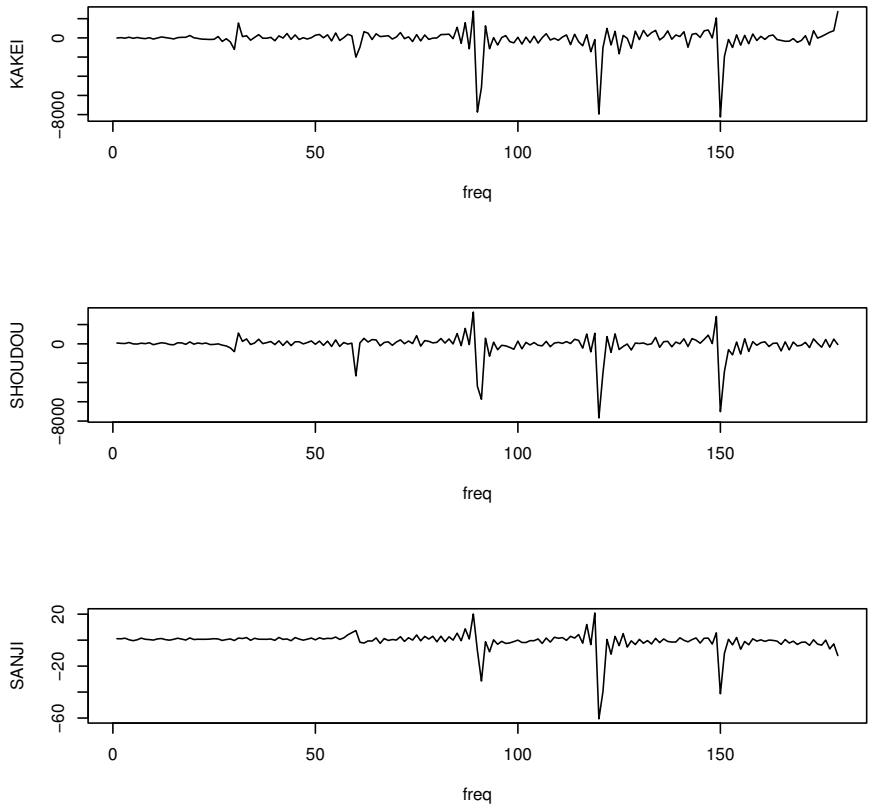


図 7.6 : 消費系列の z_n 過程

[3] Engle, R. and C.W.J. Granger (1987), "Co-integration and Error Correction," *Econometrica*, Vol.55, 251-276.

[4] Kitagawa, G. (2010), *Introduction to Time Series Analysis*, CRC Press.

[5] Johansen, S. (1995), *Likelihood Based Inference in Cointegrated Vector Autoregressive Models*, Oxford UP.

[6] Kunitomo, N. , Sato and D. Kurisu (2018), *Separating Information Maximum Likelihood Estimation for High Frequency Financial Data*, Springer.

[7] Kunitomo, N. and S. Sato (2017), "Trend, Seasonality and Economic Time Series : the Non-stationary Errors-in-variables Models," SDS-4, MIMS, Meiji University, http://www.mims.meiji.ac.jp/publications/2017-ds.

[8] Kunitomo, N., N. Awaya and D. Kurisu (2017), "Some Properties of Estimation Methods for Structural Relationships in Non-stationary Errors-in-Variables Models," SDS-3, MIMS, Meiji University.

[9] Kunitomo, N. and S. Sato (2019), " A Robust-filtering Method for Noisy Non-Stationary Time Series," Unpublished Manuscript.

[10] Müller, U. and M. Watson (2018), "Long-run Covariability," *Econometrica*, 86-3, 775-804.

[11] Nishimura, K.G. S. Sato and A. Takahashi (2019), "Term Structure Models During the Global Financial Crisis: A Parsimonious Text Mining Approach," Asia-Pasific Financial Markets.

テキストデータからの情報抽出を利用した時系列予測†

川崎能典

統計数理研究所

2019年1月31日

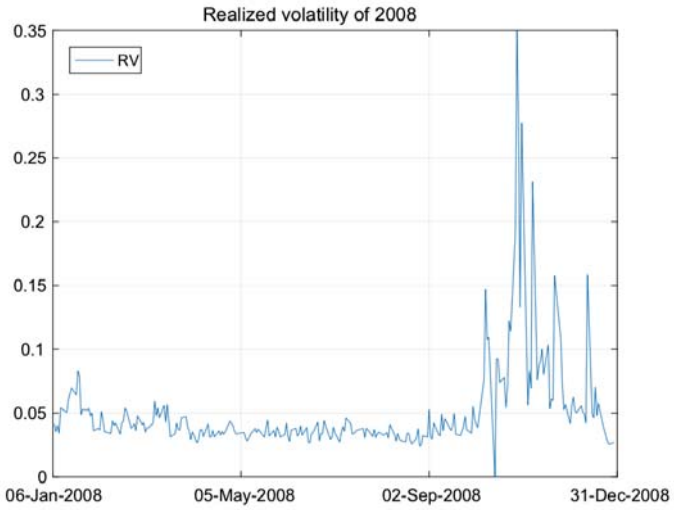
科研費基盤(A)「経済統計・政府統計の理論と応用」
研究集会＠東京大学経済学部

†森本孝之氏 (関西学院大学理工学部)との共著

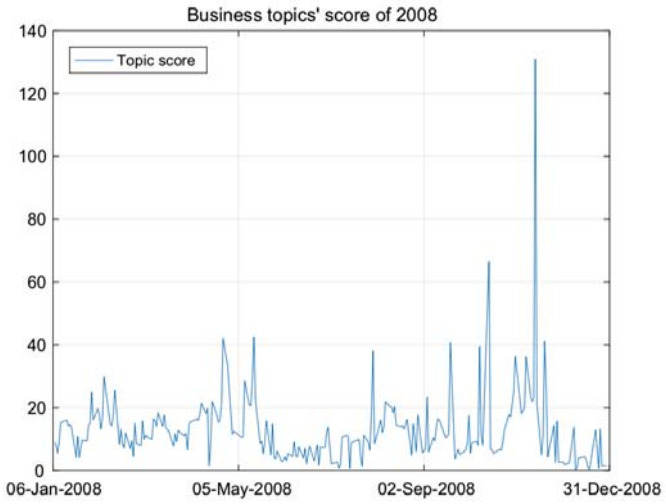
Motivation

- Counts of keywords sometimes helps
  - (Ex.) Google SVI (Search Volume Index)
- Have to find nice keywords.
- From news (text) data, we want to extract topics (defined by distribution of words) that may affect market sentiments
- Construct topic score time series SC_t
- Investigate if SC_t improves volatility forecasting

Illustration: topic score and realized volatility



Realized volatility estimated from high frequent data



Estimated topic score (one of 20 scores)

"Bag-of-Words" model

- We only focus on word frequencies, and neglect other information (order of words, dependency and so on.
- (Ex.) A document $D =$ "It is fine today" can be expressed $D = \{"it", "is", "fine", "today"\}$.
- Usually we exclude so-called "stop words" such as "a", "the", "for", etc.
- In this research, after morphological analysis, we choose nouns only, and remove numerals, suffixes, non-independent words, pronouns and symbols.

Latent Dirichlet Allocation Model

- A standard method for topic modeling
- Often abbreviated as LDA
- Distribution of words follows multinomial distribution (gives likelihood)
- Dirichlet distribution gives a prior distribution of words frequencies
- Word distribution ϕ_z characterizes a topic z, and each document d consists of many topics of which distribution θ_d.

Typical MCMC cycle for LDA

1. For $k \in \{1 \ldots Z\}$:
   - Generate a word distribution for each topic, $\phi_z \sim \text{Dirichlet}(\boldsymbol{\phi}|\boldsymbol{\beta})$,
2. For each document $d \in D$:
   - Generate a topic distribution for each document, $\theta_d \sim \text{Dirichlet}(\boldsymbol{\theta}|\boldsymbol{\alpha})$,
   - for each word $w_i \in d$:
     (a) Generate a topic from a multinomial distribution, $z_i \sim \text{Multinomial}(\theta_d)$,
     (b) Generate a word from a multinomial distribution, $w_i \sim \text{Multinomial}(\phi_{z_i})$,

where $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ denote hyper parameters of a Dirichlet distribution.

This algorithm is for a single document. We do this day by day for Reuters news, and want to ensure some continuity of topics along time axis.

Multiscale Dynamic Topic Model

- Proposed by Iwata et al. (2010)
- Parameter $\phi_{t,z}$ of word distribution of topic z at time t has some time dependent structure.

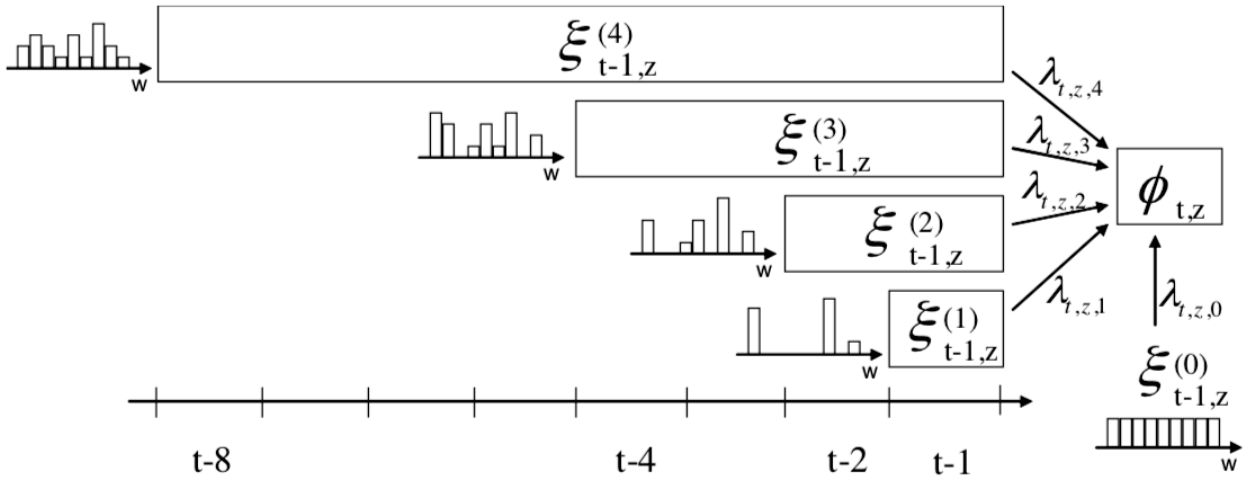
$$\phi_{t,z} \sim \text{Dirichlet}\left(\sum_{s=0}^{S} \lambda_{t,z,s}\widehat{\omega}_{t-1,z}^{(s)}\right)$$

- $\widehat{\omega}_{t-1,z}^{(s)}$: distribution of words over topic z with scale s at time $t-1$
- $\lambda_{t,z,s}$: weight for scale s in topic z at time t

Dirichlet parameter in detail

- $\widehat{\omega}_{t-1,z}^{(s)}$ indicated the word distribution (w.d.) from epoch $(t-1)-2^{s-1}+1$
- If $S=4$, s runs through $0,1,2,3,4$.
- $s=4$ →w.d. comes from $t-8$ to $t-1$
- $s=3$ →w.d. comes from $t-4$ to $t-1$
- $s=2$ →w.d. comes from $t-2$ to $t-1$
- $s=1$ →word distribution comes at $t-1$
- $s=0$; assume uniform distribution for $\widehat{\omega}_{t-1,z}^{(0)}$

Illustration of Multiscale Word Distribution



Word distributions are likely to be smoothed as the time scale becomes long.

Iwata, T. et al. (2000) Proceedings of 16th ACM SIGKDD, p.663-672.

MCMC cycle for MDTM

1. For each topic $k = 1, \ldots, Z$:
   (a) Draw word distribution of topic $\phi_{t,z} \sim \text{Dirichlet}(\sum_{s=0}^{S} \lambda_{t,z,s} \hat{\omega}_{t-1,z}^{(s)})$,
   (b) Draw a hyper parameter of prior of topic distribution $\alpha_{t,z} \sim \text{Gamma}(\zeta \alpha_{t-1,z}, \zeta)$,
2. For each document $d = 1, \ldots, D_t$:
   (a) Draw topic proportions $\theta_{t,d} \sim \text{Dirichlet}(\alpha_t)$,
   (b) For each word $n = 1, \ldots, N_{t,d}$:
      i. Draw topic $z_{t,d,n} \sim \text{Multinomial}(\theta_{t,d})$,
      ii. Draw word $w_{t,d,n} \sim \text{Multinomial}(\phi_{t,z_{t,d,n}})$,

Weights $\{\lambda_{t,z,s}\}$ and hyperparameter $\alpha_{t-1,z}$ are estimated in an outer loop of this cycle, by stochastic EM algorithm and fixed point iteration method.

Construction of topic scores

- Topic scores are made up by estimated topic proportions $\theta_{t,j,i}$ (percentage of topic i included in j-th document at time t)

$$SC_t^i = \sum_{j=d}^{D_t} \theta_{t,j,i}$$

- SC_t^i: score for topic i at time t
- D_t: number of documents at time t
- $\theta_{t,j,i}$: i-th element of the topic distribution within j-th document at time t

Word distribution (June 2, 2008)

Topic 1		Topic 2	
Nikkei	0.109	Yen	0.208
Average	0.107	Present	0.069
Continued rise	0.043	Weekend	0.053
TSE	0.038	Temporary	0.040
Center	0.037	Higher quotation	0.038
Mutual fund	0.035	Feasible	0.036
Domestic	0.032	Session	0.034
Tokyo	0.032	Rebound	0.033
Opening	0.031	Late	0.030
Major	0.027	Holidays	0.024

- We consider 20 topics in all.
- Word distribution in Topic 1 and Topic 2
- Only top 10 words are shown

Data

- High frequent data of stock index (TOPIX)
- January 7th 2008 – December 28th 2012, $T = 1223$
- Generate 1 min return and calculate daily realized volatility (RV_t) and realized quarticity (RQ_t)
- News data taken from Reuter Japan's web site
- Language = Japanese
- 298,205 documents, 24,227 non-overlapping words excluding stop words

Forecasting models

- Heterogeneous Autoregressive (HAR) model
  - ✓Baseline model, Corsi (2009)
- HARQ model, adding realized quarticity (RQ_{t-1}) in the coefficient of RV_{t-1}
  - ✓Bollerslev, Patton and Quaedvleig (2016)
- HAR + topic score (HAR-SC)
- HARQ + topic score (HARQ-SC)
  - In our paper, we did AR vs. AR-SC and ARQ vs. ARQ-SC comparison which will be omitted here.

HAR vs. HAR-SC

- HAR-SC model is defined by

$$RV_t = \beta_0 + \beta_1 RV_{t-1} + \beta_2 RV_{t-1|t-5} + \beta_3 RV_{t-1|t-22} + \gamma SC_{t-1} + u_t$$

where $RV_{t-j|t-h} = \frac{1}{h+1-j} \sum_{i=j}^{h} RV_{t-i}$

- Omitting γSC_{t-1} reduces to HAR model

HARQ vs. HARQ-SC

- HARQ-SC model is defined by

$$RV_t = \beta_0 + (\beta_1 + \beta_{1Q} RQ_{t-1}^{1/2}) RV_{t-1} + \beta_2 RV_{t-1|t-5} + \beta_3 RV_{t-1|t-22} + \gamma SC_{t-1} + u_t$$

where $RV_{t-j|t-h} = \frac{1}{h+1-j} \sum_{i=j}^{h} RV_{t-i}$

- Omitting γSC_{t-1} reduces to HARQ model

Out-of-sample forecast losses

	HAR	HARQ	HAR-SC	HARQ-SC	
MSE (RW)	1.000	0.5562	0.9658	**0.5369**	$SC^{(11)}$
MSE (IW)	1.000	0.8408	0.9678	**0.8175**	$SC^{(11)}$
QLIKE (RW)	1.000	1.3781	**0.9891**	1.3439	$SC^{(3)}$
QLIKE (IW)	1.000	1.1529	**0.9883**	1.1292	$SC^{(18)}$

MSE: $L(RV_t, X_t) = (RV_t - X_t)^2$

QLIKE: $L(RV_t, X_t) = \frac{RV_t}{X_t} - \log\left(\frac{RV_t}{X_t}\right) - 1$

IW: increasing window in regression

RW: rolling regression with window size 400 days

Discussion

- SC_{t-1} is estimated based on information up to $t-1$. So no cheat.
- We need some preliminary analysis to search promising $SC^{(j)}$.
- Forecasting performance depends on the choice of error function as well as the choice of regression window.

HAR-HARSC: another complication

- HAR-HARSC model is defined by

$$RV_t = \beta_0 + \beta_1 RV_{t-1} + \beta_2 RV_{t-1|t-5} + \beta_3 RV_{t-1|t-22} + \gamma_1 SC_{t-1} + \gamma_2 SC_{t-1|t-5} + \gamma_3 SC_{t-1|t-22} + u_t$$

where $SC_{t-j|t-h} = \frac{1}{h+1-j} \sum_{i=j}^{h} SC_{t-i}$

HARQ-HARQSC: yet another complication

- HARQ-HARQSC model is defined by

$$RV_t = \beta_0 + (\beta_1 + \beta_{1Q} RQ_{t-1}^{1/2}) RV_{t-1} + \beta_2 RV_{t-1|t-5} + \beta_3 RV_{t-1|t-22} + \gamma_1 SC_{t-1} + \gamma_2 SC_{t-1|t-5} + \gamma_3 SC_{t-1|t-22} + u_t$$

where $SC_{t-j|t-h} = \frac{1}{h+1-j} \sum_{i=j}^{h} SC_{t-i}$

Reference

- Morimoto, T. and Kawasaki, Y. (2017). Forecasting Financial Market Volatility Using a Dynamic Topic Model, *Asia-Pacific Financial Markets*, Vol. 24, pp. 149-167. DOI: 10.1007/s10690-017-9228-z
- And references therein