高頻度観測金融データに対する最尤型・ベイズ型推定法

荻原　哲平 （統計数理研究所）

1　概要

近年株式市場における全取引の情報を記録した「高頻度観測金融データ」の利用可能性が高まり, それを用いた統計解析が活発に研究されている. このようなデータはその膨大な情報量から金融市場のミクロ構造のさらなる解明が可能になると期待されるが, その特有のデータ構造から幾つかの統計解析上の問題が生じる.

まず, 証券の高頻度観測データを確率過程でモデリングする際には無視できないレベルのノイズがあることが実証研究から示唆されている. このようなノイズは「マーケット・マイクロストラクチャー・ノイズ（MMN）」と呼ばれる. また, 日内において株価の観測されるのは株価の約定時もしくは新しい注文の到着時であるため, 異なる株式に対して観測時刻が一致していないという「非同期観測」の問題が必然的に生じる.

本報告では, MMNと非同期観測の存在下におけるパラメータ付拡散過程モデルの未知パラメータの最尤型・ベイズ型推定量を構築し, その一致性, 漸近混合正規性や特殊なケースにおける漸近有効性, つまり推定量としての漸近的な最適性に関する結果を紹介する.

2　設定

以下の確率微分方程式を満たす 2 次元拡散過程 $X = \{X_t\}_{0\le t\le T}$ を考える：

$$dX_t = \mu(t, X_t, \sigma_*)dt + b(t, X_t, \sigma_*)dW_t, \quad t \in [0, T]. \tag{1}$$

ここで, $\sigma_* \in \Lambda$ はモデル・パラメータ, Λ : $\mathbb{R}^d$ 上の bounded open set, $\{W_t\}_{0\le t\le T}$: 2 次元 standard Wiener process, μ : $\mathbb{R}^2$ 値 Borel 関数, $b = \{b^{ij}\}_{i,j=1}^2$: $\mathbb{R}^2 \otimes \mathbb{R}^2$ 値 Borel 関数とする. X の成分を X^1, X^2 と書き, $n \in \mathbb{N}$ に対して, X^1, X^2 の観測時刻をそれぞれ random times $\{S_i^{n,1}\}_{i=0}^{\ell_{1,n}}, \{S_j^{n,2}\}_{j=0}^{\ell_{2,n}} \subset [0, T]$ で表す. 観測数のオーダーは n, つまり, $0 < P\text{-}\lim_{n\to\infty}(\ell_{k,n}/n) < \infty$ a.s. for $k = 1, 2$ とし, $\max_{i,p} |S_i^{n,p} - S_{i-1}^{n,p}| \to^p 0$ $(n \to \infty)$ を満たすとする（高頻度観測極限）. 観測ノイズ : $\{\epsilon_i^k\}_{k=1,2,i\in\mathbb{Z}_+}$ を (X_t, W_t) と独立で, $E[\epsilon_i^k] = 0$, $E[\epsilon_i^k \epsilon_j^l] = v_{k,*}\delta_{kl}\delta_{ij}$ を満たす random variables とする. 観測データが $Y_i^1 = X^1_{S_i^{n,1}} + \epsilon_i^1, Y_j^2 = X^2_{S_j^{n,2}} + \epsilon_j^2$ で与えられる時, モデルの未知パラメータ σ_* を推定する問題を考える.

3　最尤型推定量の構築

最尤推定量を構築するには尤度関数が必要だが, 一般の拡散過程に対する尤度関数の計算は困難であるため, 疑似尤度関数を構築する. 正数列 $\{l_n\}_n$ を $l_n \to \infty, l_n/\sqrt{n} \to 0, l_n/n^{1/4} \to \infty$ を満たすものとする. 観測区間全体 $[0, T]$ を l_n 個の等間隔の区間 $[u_0, u_1), \cdots, [u_{l_n-1}, u_{l_n})$ に分割して, それぞれの部分区間で疑似尤度関数を構築する.

それぞれの区間 $[u_{k-1}, u_k)$ において X^j の増分 $X_t^j - X_s^j$ は $X_t^j - X_s^j = b_k^j(\sigma_*) \cdot (W_t - W_s)$ と近似される. ただし, $b_k^j(\sigma)$ は $[u_{k-1}, u_k)$ における $b^{j\cdot}(t, X_t, \sigma_*)$ の可予測・観測可能な近似量であり, drift 項は diffusion 項よりもゼロへの収束が速いため無視できることを用いた. ノイズ ϵ_i^k が正規分布に従うとき, $[u_{k-1}, u_k)$ における観測の増分列

$$Z_k = \left((Y_i^1 - Y_{i-1}^1)^\top_{i;S_{i-1}^{n,1}, S_i^{n,1} \in [u_{k-1}, u_k)}, (Y_j^2 - Y_{j-1}^2)^\top_{j;S_{j-1}^{n,2}, S_j^{n,2} \in [u_{k-1}, u_k)} \right)$$

の分布は局所的に正規分布で近似され, その（条件付）分散共分散行列は

$$\begin{aligned} E[(Y_i^p - Y_{i-1}^p)(Y_j^p - Y_{j-1}^p)|\mathcal{F}_{u_{k-1}}] &\approx (L_{p,k}(\sigma_*) + M_{p,k})_{ij}, \\ E[(Y_i^1 - Y_{i-1}^1)(Y_j^2 - Y_{j-1}^2)|\mathcal{F}_{u_{k-1}}] &\approx (L_{1,2,k}(\sigma_*))_{ij} \end{aligned}$$

と近似される. ただし, $(L_{p,k})_{ij} = |b_k^p(\sigma_*)|^2(S_i^{n,p} - S_{i-1}^{n,p})\delta_{ij}$, $(L_{1,2,k})_{i,j} = b_k^1 \cdot b_k^2(S_i^{n,1} \wedge S_j^{n,2} - S_{i-1}^{n,1} \vee S_{j-1}^{n,2})_+$, $(M_{p,k})_{i,j} = 2\delta_{ij} - \delta_{\{|i-j|=1\}}$.

よって,

$$S_k(\sigma, v_1, v_2) = \begin{pmatrix} L_{1,k}(\sigma) + v_1 M_{1,k} & L_{1,2,k}(\sigma) \\ L_{1,2,k}(\sigma) & L_{2,k}(\sigma) + v_2 M_{2,k} \end{pmatrix}$$

に対して, 疑似対数尤度関数 $H_n(\sigma, v_1, v_2)$ を

$$H_n(\sigma, v_1, v_2) = -\frac{1}{2}\sum_{k=2}^{l_n}\{Z_k^\top S_k^{-1}(\sigma, v_1, v_2)Z_k + \log\det S_k(\sigma, v_1, v_2)\}$$

と定める. ϵ_i^k が正規分布に従わないときは H_n は尤度関数の近似になっていないが, このようなノイズの分布の misspecification は σ_* の推定に影響を与えないことが示される. よって非正規のノイズに対してもこの H_n を用いて推定量を構築する.

個々の観測の変動において観測ノイズの影響が支配的であるため, 観測ノイズの分散 $v_{1,*}, v_{2,*}$ の一致推定量 $\hat{v}_{1,n}, \hat{v}_{2,n}$ の構築は容易である. 例えば $\hat{v}_{p,n} = \sum_i (Y_i^p - Y_{i-1}^p)^2/(2\ell_{p,n})$ とすればよい. この時未知パラメータ σ_* の最尤型推定量 $\hat{\sigma}_n$ は

$$\hat{\sigma}_n = \mathrm{argmax}_\sigma H_n(\sigma, \hat{v}_{1,n}, \hat{v}_{2,n})$$

と定義される.

4 主定理

$r_n = \max_{p,i}(S_i^{n,p} - S_{i-1}^{n,p})$, $\underline{r}_n = \min_{p,i}(S_i^{n,p} - S_{i-1}^{n,p})$ と置き, 以下を仮定する.

*A*1. μ は局所有界, $b(t,x,\sigma)$ は (t,x,σ) に関して滑らか, かつ $bb^\top$ は任意の (t,x,σ) に対して正定値

*A*2. $\inf_{\sigma_1 \neq \sigma_2}(|bb^\top(t,x,\sigma_1) - bb^\top(t,x,\sigma_2)|/|\sigma_1 - \sigma_2|) > 0$ for any (t,x)

*A*3. 任意の $\delta > 0$ に対し, $r_n = O_p(n^{-1+\delta})$ かつ $\underline{r}_n^{-1} = O_p(n^{1+\delta})$

*A*4. ある $\eta \in (0, 1/2)$ と正値確率過程 $\{a_0^p(t)\}_{0 \le t \le T, p=1,2}$ があって, $a_0^p(t)$ は t に関し C^1 で $n \to \infty$ の時,

$$l_n^{-1}\sqrt{n}\max_{1 \le l \le L_n}\left| n^{-1}(s''_{n,l} - s'_{n,l})^{-1}\#\{i; [S_{i-1}^{n,p}, S_i^{n,p}) \subset [s'_{n,l}, s''_{n,l})\} - a_0^p(s'_{n,l})\right| \to^p 0.$$

ただし, $\{[s'_{n,l}, s''_{n,l})\}_l \subset [0,T]$ は任意の disjoint intervals ($n \in \mathbb{N}$) で

$$0 < \inf_{n,l}(n^{1-\eta}(s''_{n,l} - s'_{n,l})) \le \sup_{n,l}(n^{1-\eta}(s''_{n,l} - s'_{n,l})) < \infty$$

を満たすものとする

[*A*4] は直観的には任意の局所区間における大数の法則を表している.

Theorem 4.1. [*A*1]-[*A*4] を仮定する. この時, $\Gamma = P\text{-}\lim_{n\to\infty}(-n^{-1/2}\partial_\sigma^2 H_n(\sigma_*, v_{1,*}, v_{2,*}))$ は非退化行列であり, ある $\zeta \sim N(0, I_d) \perp \Gamma$ があって,

$$n^{1/4}(\hat{\sigma}_n - \sigma_*) \to^d \Gamma^{-1/2}\zeta, \quad -n^{-1/2}\partial_\sigma^2 H_n(\hat{\sigma}_n, \hat{v}_{1,n}, \hat{v}_{2,n}) \to^p \Gamma \quad \text{as} \quad n \to \infty.$$

特に

$$(-\partial_\sigma^2 H_n(\hat{\sigma}_n, \hat{v}_{1,n}, \hat{v}_{2,n}))^{1/2}(\hat{\sigma}_n - \sigma_*) \to^d N(0, I_d) \quad \text{as} \quad n \to \infty.$$

最後の収束から最尤型推定量の信頼区間が漸近的に計算される.

また, 最尤型推定量の最適性に関する以下の結果が成り立つ.

Theorem 4.2. [*A*1]-[*A*4] を仮定する. さらに $\mu \equiv 0$, $b(t,x,\sigma)$ が (t,x) に依らず, 観測ノイズが正規分布に従うとする. この時, 統計モデルの局所漸近正規性が成り立ち, 最尤型推定量 $\hat{\sigma}_n$ は漸近有効となる.

Table 1: 最尤型推定量の simulation 結果 1

		σ_1	σ_2	σ_3	v_1	v_2
	true values	1	0.866	0.5	0.001	0.001
$n=1000$	$\hat{\sigma}_n$	0.899	0.783	0.456	0.00150	0.00151
		(0.041)	(0.041)	(0.058)	(0.00008)	(0.00007)
	$\hat{\sigma}'_n$	0.972	0.849	0.493	0.00109	0.00110
		(0.047)	(0.047)	(0.061)	(0.00007)	(0.00007)
	$\hat{\sigma}''_n$	0.999	0.873	0.507	-	-
		(0.046)	(0.046)	(0.061)	-	-
$n=5000$	$\hat{\sigma}_n$	0.967	0.831	0.484	0.00110	0.00110
		(0.027)	(0.030)	(0.042)	(0.00003)	(0.00003)
	$\hat{\sigma}'_n$	1.000	0.859	0.499	0.00101	0.00101
		(0.029)	(0.032)	(0.044)	(0.00003)	(0.00003)
	$\hat{\sigma}''_n$	1.004	0.862	0.501	-	-
		(0.029)	(0.030)	(0.044)	-	-

疑似対数尤度関数 H_n を用いてベイズ型推定量を構築することも可能である. 事前確率密度 $\pi(\sigma)$ は有界連続関数で $\inf_\sigma \pi(\sigma) > 0$ を満たすとする. この時二次損失関数に対するベイズ型推定量 $\tilde{\sigma}_n$ は以下のように定義できる.

$$\tilde{\sigma}_n = \left(\int \exp(H_n(\sigma, \hat{v}_n))\pi(\sigma)d\sigma \right)^{-1} \int \sigma \exp(H_n(\sigma, \hat{v}_n))\pi(\sigma)d\sigma$$

Theorem 4.3. 幾つかの [A1]-[A4] より強い仮定を置く *(μ_t, $b(t, X_t, \sigma)$ や [A4] の収束列に対するモーメント条件など). この時, Theorem 4.1* の Γ と ζ と, 任意の有界確率変数 $\mathbf{Y}$ と任意の高々多項式増大の連続関数 f に対して

$$\begin{aligned} E[\mathbf{Y}f(n^{1/4}(\hat{\sigma}_n - \sigma_*))] &\to E[\mathbf{Y}f(\Gamma^{-1/2}\zeta)], \\ E[\mathbf{Y}f(n^{1/4}(\tilde{\sigma}_n - \sigma_*))] &\to E[\mathbf{Y}f(\Gamma^{-1/2}\zeta)]. \end{aligned}$$

5 シミュレーション

シンプルなモデル:

$$dY_t^1 = \sigma_1 dW_t^1, \quad dY_t^2 = \sigma_3 dW_t^1 + \sigma_2 dW_t^2$$

に対し, 最尤型推定量 $\hat{\sigma}_n$ を計算する. ただし $\sigma = (\sigma_1, \sigma_2, \sigma_3)$ をパラメータとして $Y_0 = 0$ とする. 観測は平均パラメータ $\lambda_1, \lambda_2 > 0$ の独立ポアソン過程 $\{\bar{N}_t^1\}, \{\bar{N}_t^2\}$ に対して, $S_i^{n,p} = \inf\{t \geq 0; \bar{N}_{nt}^p \geq i\}$ で与えられるとする. 観測ノイズ $\{\epsilon_i^{n,p}\}_i \sim^{i.i.d.} N(0, v_{p,*})$ に対し, 観測は $Y_i^p = X^p_{S_i^{n,p}} + \epsilon_i^{n,p}$ で与えられる. $k_n = [n^{5/8}]$ を使用し, $\hat{v}_{p,n} = \sum_i (\Delta Y_i^p)^2/(2\ell_{p,n})$ とし, 対応する最尤型推定量を $\hat{\sigma}_n$ とする. 最尤型推定量 $\hat{\sigma}_n$ をプラグインした $\hat{v}'_{p,n} = \hat{v}_{p,n} - |b^p(\hat{\sigma}_n)|^2 T/(2\ell_{p,n})$ も考え, 対応する推定量を $\hat{\sigma}'_n$ とおく. また $H_n(\sigma, v)$ に真値 v_* を代入して計算した推定量を $\hat{\sigma}''_n$ とおく. Table 1 は $T = 1, (\lambda_1, \lambda_2) = (1, 1)$, $(\sigma_{1,*}, \sigma_{2,*}, \sigma_{3,*}) = (1, \sqrt{1 - 0.5^2}, 0.5)$, $v_* = (0.001, 0.001)$ として 1000 回シミュレーションをした時の推定量の標本平均と標本標準偏差（括弧内）を表している. 最尤型推定量をプラグインする前の v の推定量は上方のバイアスがあり, σ の推定量は下方のバイアスがあることがわかる. これは推定量 $\hat{v}_{p,n}$ には常に正となる潜在確率過程 X の二次変分を項として含んでいるためと考えられる. v が既知の場合の推定量である $\hat{\sigma}''_n$ とプラグインして v を推定した $\hat{\sigma}'_n$ は $n = 5000$ ではさほどパフォーマンスが変わらず, v の推定量として $\hat{v}'_n$ が妥当であることを示している. Table 2 は $v_* = (0.005, 0.005)$ と比較的大きいノイズ分散を用いた場合の結果を表している. 実際の市場におけるノイズ分散の値は潜在過程の二次変分を 1 とした場合, 概ね 0.001 から 0.005 程度の水準となっている. Table 2 においてもプラグインした推定量 $\hat{\sigma}'_n$ のパフォーマンスはサンプル数が大きいときに $\hat{\sigma}''_n$ のパフォーマンスに近いことが分かる.

また, 最尤型推定量を用いて, 二次変分 $\langle X^1, X^2 \rangle_T$ の推定量 $\hat{\sigma}_{1,n}\hat{\sigma}_{3,n}T$ も考えることができ, パラメータの付け替えに依って $\sigma_1\sigma_3$ をパラメータにすることができるので, このモデルにおいては $\hat{\sigma}_{1,n}\hat{\sigma}_{3,n}T$ も漸近有効である. この推定量のパフォーマンスを既存のノンパラメトリック・セミパラメトリック推定量と比較する. 比較する推定量として, Christensen, Kinnebrock and Podolskij (2010) の pre-averaged HY (PHY), Modulated Realized Covariance (MRC), Bibinger et al. (2014) の local method of moments (LMM), Bibinger (2011) の Generalized Multiscale Estimator(GME) を用いる. これらの推定量は LMM 以外は R パッケージ"yuima"の"cce"関数を用いて計算することができる. LMM は Bibinger et al. (2014) の oracle estimator を用いる. これは一部真値の情報を用いており, 真値の情報を用いない推定量も提案されているがより複雑で推定誤差は大きい.

Table 2: 最尤型推定量の simulation 結果 2

		σ_1	σ_2	σ_3	v_1	v_2
	true values	1	0.866	0.5	0.005	0.005
$n = 1000$	$\hat{\sigma}_n$	0.955	0.834	0.472	0.00546	0.00551
		(0.064)	(0.061)	(0.090)	(0.00032)	(0.00032)
	$\hat{\sigma}'_n$	0.990	0.865	0.488	0.00500	0.00505
		(0.069)	(0.066)	(0.095)	(0.00032)	(0.00033)
	$\hat{\sigma}''_n$	0.991	0.868	0.490	-	-
		(0.070)	(0.067)	(0.094)	-	-
$n = 5000$	$\hat{\sigma}_n$	0.983	0.855	0.488	0.00509	0.00511
		(0.048)	(0.038)	(0.059)	(0.00012)	(0.00011)
	$\hat{\sigma}'_n$	0.992	0.863	0.493	0.00500	0.00501
		(0.049)	(0.038)	(0.060)	(0.00012)	(0.00011)
	$\hat{\sigma}''_n$	0.992	0.863	0.493	-	-
		(0.048)	(0.038)	(0.059)	-	-

Table 3: 二次変分 $\langle X^1, X^2 \rangle_T$ の推定量の比較

	n	最尤型	PHY	GME	MRC	MRC2	LMM
$v_1 = 0.001$	500	0.458	0.522	0.511	0.538	0.522	0.391
$v_2 = 0.001$		(0.078)	(0.143)	(0.142)	(0.257)	(0.129)	(0.079)
	1000	0.480	0.502	0.490	0.485	0.502	0.468
		(0.070)	(0.133)	(0.129)	(0.180)	(0.118)	(0.076)
	5000	0.500	0.504	0.503	0.502	0.506	0.505
		(0.049)	(0.088)	(0.077)	(0.131)	(0.079)	(0.073)
$v_1 = 0.005$	500	0.499	0.499	0.497	0.524	0.528	0.442
$v_2 = 0.005$		(0.137)	(0.176)	(0.190)	(0.242)	(0.175)	(0.126)
	1000	0.485	0.478	0.466	0.468	0.483	0.507
		(0.108)	(0.163)	(0.137)	(0.192)	(0.122)	(0.118)
	5000	0.490	0.488	0.489	0.505	0.492	0.506
		(0.072)	(0.103)	(0.095)	(0.125)	(0.086)	(0.081)

Table 3 はシミュレーション結果を表している．各パラメータの値は上と同様にし, この時二次変分の真値は $\langle X^1, X^2 \rangle_T = \sigma_{1,*}\sigma_{3,*}T = 0.5$ となる．最尤型推定量は $\hat{\sigma}'_n$ を採用し, 他の推定量のチューニングパラメータは MRC2 以外は"cce"関数のデフォルト値を使用した．また MRC2 は $\theta = 1/3$ を使用した (PHY は $\theta = 0.15$, MRC は $\theta = 1$, LMM は $J = 30$, $h^{-1} = 10$)．どの推定量においてもサンプル数の増大につれて標本平均が真値に近づいているが, 標本標準偏差において最尤型推定量が最も小さい値を取っていることが分かる．