2017年6月16日 統計数理研究所 オープンハウス

高頻度金融データのモデリングと統計解析

荻原 哲平　　数理・推論研究系 助教

1. 高頻度金融データ

株式市場の株価データからその変動の大きさや異なる株式間での株価の連動性などの特性を統計的に解析することは、株式資産のリスクを管理する上で重要な課題と言える. 従来の金融資産のリスク管理においては, 一日の終わりの株価を用いるなどの日単位の株価データを用いた分析が主流だったが, 近年は一日内の全ての株式取引のデータに関する情報がデータベースで管理されるようになり, それらを用いた研究が活発になっている. それらは秒・ミリ秒単位の高頻度に観測されたデータであり, 日単位のデータに比べて非常に多くの情報を含んでいる. このような高頻度金融データのモデリングを考える.

2. 拡散過程による株価のパラメトリック・モデル

【高頻度金融データ特有の問題】

高頻度金融データは, そのデータの膨大さに加え, 特有の複雑な構造を持つため統計解析は容易ではない. ここでは拡散過程による株価モデリングを考える際の重要な問題として以下の二点を取り上げる.

i.　マーケット・マイクロストラクチャー・ノイズ

高頻度金融データにおける観測ノイズの混入が実証研究より示唆されている

ii.　非同期観測

異なる株式に対する観測時刻が異なる → 従来の統計解析手法の適用が困難

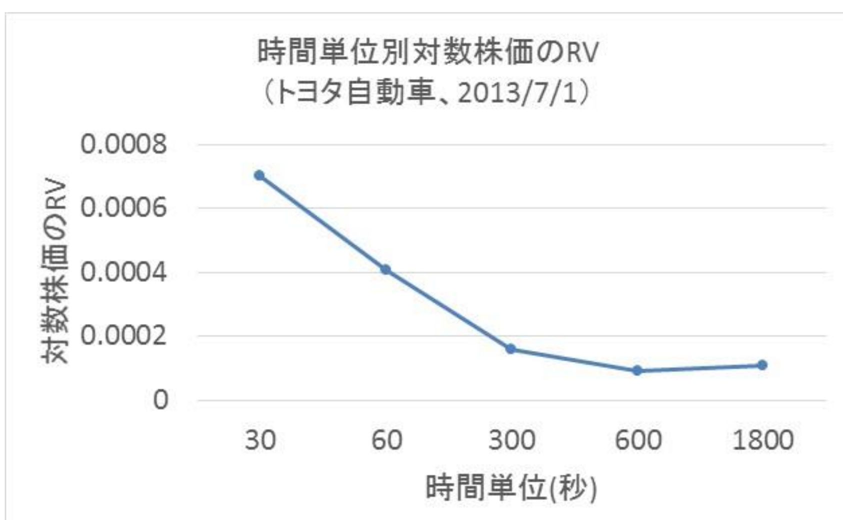
マーケット・マイクロストラクチャー・ノイズによる実現ボラティリティのバイアス

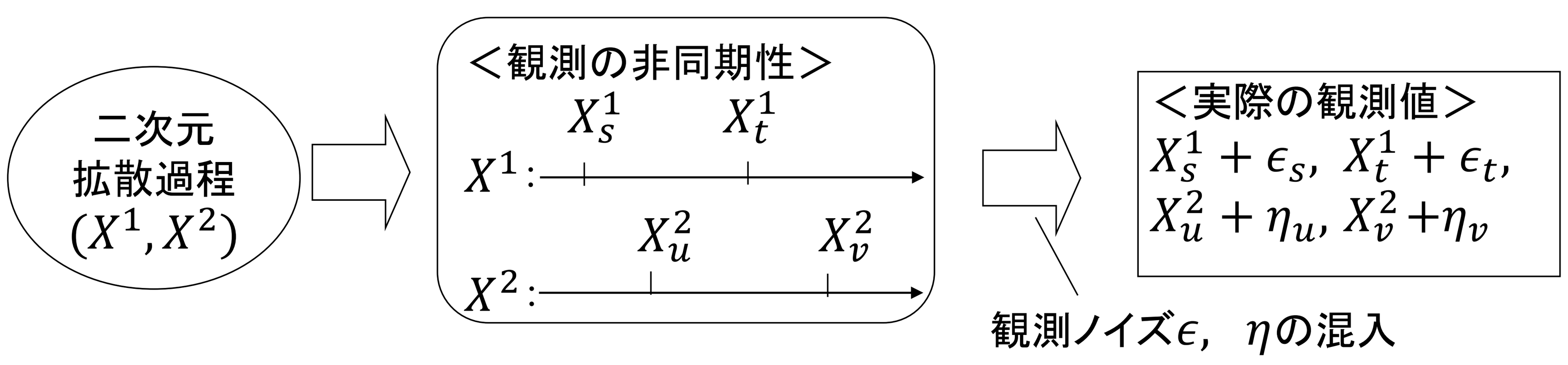
対数株価X_tが確率過程で$\{X_{t_k}\}_{k=0}^L$が観測されるとすると, $\max(t_k - t_{k-1}) \to^p 0$ (高頻度観測極限)の時に

$$\mathrm{RV}_n := \sum_{k=1}^{L}\left(X_{t_k} - X_{t_{k-1}}\right)^2 \to^p \langle X\rangle_T$$

Xの変動を表すリスク量

一方で実証研究においては右図のように高頻度になっても収束しないため, 仮想的な観測ノイズの存在が示唆されている





【拡散過程によるモデリング】

二次元の潜在株価過程 $X_t = (X_t^1, X_t^2)$ は確率微分方程式

$$dX_t = \mu(t, X_t, \sigma_*)dt + b_\circ(t, X_t)dW_t, \qquad t \in [0, T]$$

を満たすとする. ただし, W_tは標準ブラウン運動, $\mu, b_\circ$は既知の関数, σ_*はd次元未知パラメータである. X_t^1, X_t^2 の観測時刻はそれぞれ$s_0^1, s_1^1, \ldots, s_l^1$と$s_0^2, s_1^2, \ldots, s_m^2$で与えられるとする（非同期観測）. さらに平均0の独立同分布の観測ノイズ$(\epsilon_i)_{0\le i\le l}$, $(\eta_j)_{0\le j\le m}$に対して, 観測データは

$$Y_i^1 = X_{s_i^1}^1 + \epsilon_i \ (0 \le i \le l), \qquad Y_j^2 = X_{s_j^2}^2 + \eta_j \ (0 \le j \le m)$$

で与えられるとする. 観測データから$b_\circ(t, x)$の関数形を推定することで株価過程のボラティリティや共変動などのリスク量を計算することができる.

【最尤型推定法】

パラメトリック・モデル$\{b(t, x, \sigma)\}_\sigma$を考え, あるパラメータの値$\sigma_*$に対し $b_\circ(t, x) \equiv b(t, x, \sigma_*)$とする. この時, 潜在株価過程$X_t$の局所ガウス近似を用いることにより$b(\sigma) = \{b(t, X_t, \sigma)\}_{0\le t\le T}$に対して尤度関数の近似$H_n(b(\sigma))$が得られ, 最尤型推定量$\hat{\sigma}_n$は$\hat{\sigma}_n = \mathrm{argmax}_\sigma H_n(b(\sigma))$により得られる. 観測数のオーダーを$b_n$と書くと, 提案推定量$\hat{\sigma}_n$は一定の条件の下$n \to \infty$で以下の漸近混合正規性を満たす:

$$b_n^{1/4}(\hat{\sigma}_n - \sigma_*) \to^d \Gamma^{-1/2}\zeta.$$

ただし, Γはある確率行列で, ζはΓと独立な多変量標準正規乱数である.

2. 機械学習を用いた株価モデリング

【高頻度金融データのモデルの複雑さ】

高頻度金融データには上にあげた観測の複雑さに加えてモデル自体にもいくつかの特徴があり、シンプルなモデルによる記述が困難になっている.

・日内・曜日季節性

取引開始・終了付近へ取引が集中する等, 一日・曜日単位の周期性が存在

・長期記憶性

市場にショックがあった時にボラティリティが上がった状態が長く続く

・分布の裾での相関の高まり

・複雑な高次元の相関構造（業種・企業規模・財務指標）

etc.

- 個々の問題は研究されているが, これらの問題を統一的に扱うことのできる具体的なモデルは提案されていない
  →機械学習理論を用いて、データから構造を学習することを考えていく.

【機械学習理論の適用】

機械学習では通常真のモデルがパラメトリック・モデルに含まれることを仮定しない, つまり, 拡散係数のパラメトリック・モデル$\{b(t, x; \beta)\}_\beta$に対して, 「あるβ_*が一意に存在して$b(t, X_t) = b(t, X_t; \beta_*)$」が成り立つとは限らない状況を考える. このようなモデルはmisspecified modelと呼ばれ, 真のモデルがパラメトリック・モデルに含まれる場合と異なった漸近的性質が現れることがある. $b(\beta) = \{b(t, X_t, \beta)\}_{0\le t\le T}$に対して

$$\hat{\beta}_n = \mathrm{argmax}_\beta H_n(b(\beta))$$

としてパラメータ推定量を定める.

【漸近的性質】

$b_\circ = \{b_\circ(t, X_t)\}_{0\le t\le T}$に対して,

$$d(b(\beta), b_\circ) := \lim_{n\to\infty}\left(b_n^{-1/2}\left(H_n(b_\circ) - H_n(b(\beta))\right)\right)$$

とおくと, $d(b(\beta), b_\circ)$は擬似対数尤度比の極限であり, 常に非負.

この時学習されたパラメータ$\hat{\beta}_n$に対して以下が成立.

定理.　$d(b(\hat{\beta}_n), b_\circ) \to^p \min_\beta d(b(\beta), b_\circ)$ as $n \to \infty$.

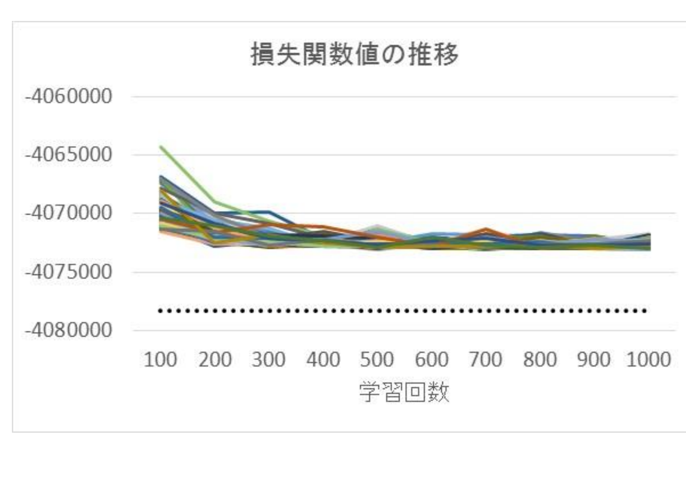
つまり, 推定量$\hat{\beta}_n$はdの意味で$b_\circ$に最も近い関数に収束する. また, 一定の条件の下,

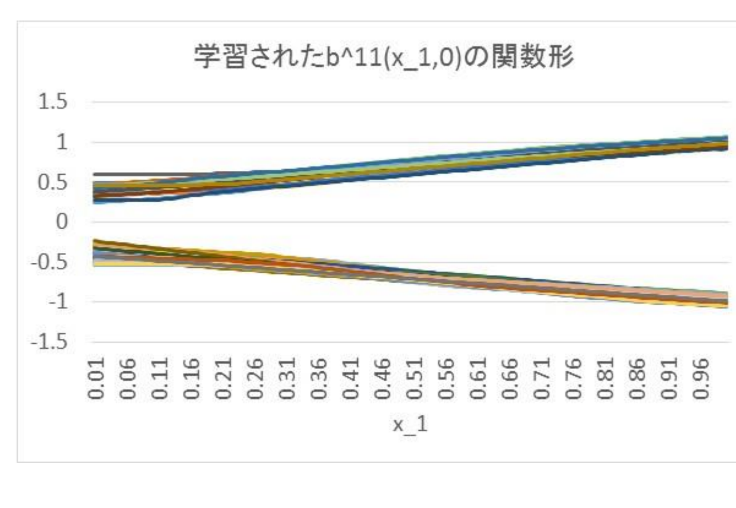
$$C_1\int_0^T |bb^\top(t, X_t, \beta) - b_\circ b_\circ^\top(t, X_t)|^2 dt \le d(b(\beta), b_\circ) \le C_2\int_0^T |bb^\top(t, X_t, \beta) - b_\circ b_\circ^\top(t, X_t)|^2 dt.$$

【シミュレーション分析】

潜在株価過程 $X_t = (X_t^1, X_t^2)$が以下を満たす時の学習結果は以下のようになった.

$$dX_t^1 = (1 - X_t^1)dt + \sigma_1\sqrt{X_t^1}dW_t^1,$$
$$dX_t^2 = (1 - X_t^2)dt + \sqrt{X_t^2}(\sigma_2 dW_t^1 + \sigma_3 dW_t^2)$$





【応用例】

高頻度金融データに対する学習理論により, 証券市場の様々な解析の進展が期待される.

・高頻度金融データを用いた株式ポートフォリオのリスク・コントロール

→高頻度金融データを用いた株価分散共分散行列の予測により, 特に急激に市場が変化する局面における適切なコントロール手法の確立

・株式間の連動性の研究

分布の裾での相関の高まりといった, 相関係数だけではとらえられない株式間の連動性を, 確率過程によるミクロ構造のモデリングでより柔軟に表現することが可能

・金融資産の売買注文の成立のしやすさ（流動性）の予測分析

→株式板情報の点過程モデリングへ応用することで, 株式売買にかかるコストの制御