高频金融数据的高维积分波动率矩阵估计

穆燕¹, 苑慧玲¹, 周勇^{1,2*}

1. 上海财经大学统计与管理学院, 上海 200433;

2. 中国科学院数学与系统科学研究院, 北京 100190

E-mail: muyan19881003@163.com, yuanhuiling2013@163.com, yzhou@amss.ac.cn

收稿日期: 2016-03-01; 接受日期: 2017-11-14; 网络出版日期: 2018-01-25; * 通信作者

国家自然科学基金 (批准号: 71331006 和 91546202)、中国科学院重点实验室 (批准号: 2008DP173182)、国家数学与交叉科学中心 (批准号: 2008DP173182) 和上海财经大学创新团队支持计划 (批准号: IRTSHUFE13122402) 资助项目

摘要 高维积分波动率矩阵是资源配置和风险管理的重要统计量, 对其估计是金融统计和风险度量中的热点和核心问题之一. 本文在带有市场信息的微观结构噪声下, 考虑了高频金融数据大量资产的积分波动率矩阵估计问题. 在多资产价格观察不同步下, 当资产数和样本量都趋向于无穷时, 本文利用不重叠区间方法和稀疏性特征提出了高维积分波动率矩阵估计, 证明了该估计量具有相合性, 较在加性噪声下的估计具有更快的收敛速度, 其收敛速度可以达到已存在高维积分波动率矩阵估计在无噪声下的最快收敛速度. 对所提出的估计与现有的高维积分波动率矩阵估计进行模拟比较, 结果表明本文提出的估计方法具有优良的性质. 最后将提出的估计应用于上海证券指数数据的实证研究中.

关键词 高频数据 高维波动率矩阵 异步性 市场微观结构 噪声

MSC (2010) 主题分类 60B12, 91G70

1 引言

高频数据和超高频数据 (日内以小时、分钟、秒或分笔为采集频率) 分析是金融计量新兴领域的最前沿热点问题之一. 自 20 世纪 90 年代, 随着信息技术的不断革新、电子交易平台在国际证券市场中的逐步推广使用, 金融市场的资源配置效率得到了提高, 人们也能方便地从市场上获取和存储 (超) 高频交易数据. 最典型的金融高频交易数据是分笔数据 (transaction by transaction data) 和分时数据 (tick by tick data). 这些海量的交易信息将较以往的低频数据 (以日、周或月为采集频率) 更能帮助我们认识市场本质, 改善市场机制设计, 进而有效地度量金融风险 and 实现市场的资源配置效率及促进金融理论的发展.

波动率矩阵作为金融资产价格变化的一种统计度量, 是资产配置、现代资产定价和金融风险管理等重要金融理论和活动的基础与核心要素. 早在 1952 年 Markowitz 的资产组合选择模型中, 波动率矩

英文引用格式: Mu Y, Yuan H L, Zhou Y. High-dimensional integrated volatility matrix estimation for high-frequency financial data (in Chinese). Sci Sin Math, 2018, 48: 319–344, doi: 10.1360/N012016-00047

阵就扮演了十分重要的角色, 对其估计的好坏直接影响着资产组合的风险. 然而大量的实证研究表明, 金融资产的波动率具有较强的记忆性和时变性, 不易被观察, 因此利用资产的积分波动率估计代替对资产点波动率准确估计和预测有利于制定更为合理的投资决策.

Andersen 等^[1] 最早提出已实现波动率 (RV, realized volatility) 作为积分波动率的一个估计, 该估计是利用高频日内数据进行直接估计, 没有假设参数模型, 不需要对参数进行估计, 克服了由于多个资产导致的传统多元广义自回归条件异方差模型 (GARCH) 的“维数灾难”. 同时高频数据比低频数据包含了更多的有用市场信息, 使用高频信息能提高估计的精度. 因此, 已实现波动率成为目前高频数据中应用最广泛的一种积分波动率估计方法.

随着研究的深入, 有些学者发现, 当考察高频观测数据受到市场微观结构噪声的影响时, Andersen 等^[1] 提出的已实现波动率 RV 并不是积分波动率的相合估计, 因此在离散数据下提出对已实现波动率估计并提高其精度和效率是重要的研究问题. 目前, Andersen 等^[1] 的已实现波动率 RV 依赖于资产市场价格的准确观察, 已有相关理论和实际应用表明资产的市场价格难以被准确观察, 通常会受到微观结构噪声的影响. 因此, 研究市场结构噪声下的积分波动率估计问题具有重要的意义. 目前最简单的噪声模型是加性噪声 (additive noise) 模型, 具体形式如下:

$$Y_{t_k} = X_{t_k} + \varepsilon_{t_k}, \quad k = 0, 1, \dots, n,$$

其中 $0 = t_0 < t_1 < \dots < t_n = T$, X_{t_k} 和 ε_{t_k} 分别记为观察时刻 t_k 的真实对数价格和市场微观结构噪声, 并且假设 ε_{t_k} 与 X_{t_k} 是相互独立的时间序列. Zhang 等^[2] 利用高频数据在此模型下首次提出了相合的积分波动率估计 TSRV (two-time scale realized volatility); Zhang^[3] 对 TSRV 进一步推广, 提出了具有更快收敛速度的相合积分波动率估计 MSRV (multiple-time scale realized volatility). 另外, 关于相合积分波动率的估计还有 WRV (wavelet realized volatility^[4])、FRV (Fourier realized volatility^[5])、RK (realized kernels^[6])、PAV (pre-averaging estimator^[7]) 和 QMLE (quasi-maximum likelihood estimation^[8]) 等.

上述都是针对单个或两个资产积分波动率的估计方法. 事实上, 在高频数据下, 多资产的积分波动率矩阵估计通常会面临异步性问题, 即不同资产交易发生的时间不同, 资产的高频价格也将被记录在不同的时间点上. Hayashi 和 Yoshida^[9] 基于重叠区间提出了新的波动率估计量, 该估计量与资产是否同步无关; Zhang^[10] 利用 PT (previous ticks) 思想提出了两资产的积分波动率矩阵估计方法; Christensen 等^[11] 结合 Hayashi 和 Yoshida^[9] 处理非同步数据方法和 PA (pre-averaging) 方法对两资产协方差阵进行估计; Ait-Sahalia 等^[12] 提出了一般的同步方法并利用 QMLE 方法对两资产的积分波动率矩阵进行了估计; Barndorff-Nielsen 等^[13] 提出了更新时间算法, 利用 RK 方法估计了多资产的积分波动率矩阵.

然而现实生活中的金融资产组合中往往会涉及几十甚至上百个资产, 相应的积分波动率矩阵也将是高维矩阵. 有关高维矩阵估计的研究一直是一个困难问题, 尤其当资产数大于样本量、已实现波动率矩阵是奇异时, 现有的估计方法不能较好地估计总体 (population) 波动率矩阵. 另外, 随着高频交易数据的记录与存储, 大量的高频数据也对估计高维波动率矩阵带来困难. 因此, 估计高维波动率矩阵成为高频金融数据研究的重点和难点.

Wang 和 Zou^[14] 及 Tao 等^[15] 对大量资产的积分波动率矩阵进行估计时发现已存在的波动率矩阵估计量具有较差的统计性质. 事实上, 当资产数和样本量趋于无穷时, 高维积分波动率矩阵的估计量是不相合的. 高维矩阵估计具有相合性通常要求估计的矩阵具有某种稀疏性, 该稀疏性可能本质上是关于高维数据中的某些低维结构的一种公式化表述. 例如, 在高频金融数据高维波动率矩阵的估计中,

稀疏作用就是保留在驱动波动率运动和捕捉市场风险中起到主导性作用的相对少数的市场因子. 当资产数和样本量趋向于无穷时, Wang 和 Zou^[14] 在假设真实积分波动率矩阵具有稀疏性的条件下, 首次提出了高维积分波动率矩阵的相合估计 ARVM (averaging realized volatility matrix); 并且证明了该估计的最大范数收敛速度达到样本量 $n^{-1/6}$. Tao 等^[16] 对 ARVM 进行了改进, 提出了收敛速度更快的高维积分波动率矩阵的相合估计 MSRV (multi-scale realized volatility matrix), 其收敛速度为 $n^{-1/4}$.

无论是单资产还是多资产, 上述的积分波动率估计方法都在简单的加性微观噪声模型下进行研究的. 然而, 资产真实价格一般市场上不易准确观察到, 为此利用资产的可观察价格视为在简单加性噪声环境下的真实价格. 这种简单的假设忽略了金融市场上重要的微观结构信息, 而这些信息包括成交量、买卖价差、交易类型和交易率等, 一般对市场价格具有重要的影响, 对提高资产价格波动率矩阵估计的精度有显著的作用. 为此, 本文将考虑带有市场信息的微观结构噪声, 利用市场上这些额外的信息提高积分波动率矩阵估计的精度. 此外, 在其他条件不变的情形下通过加入市场辅助信息来构建积分波动率矩阵估计比直接考虑与市场无关的加性噪声的估计具有更高的收敛速度.

关于市场微观结构噪声的研究可以追溯到 20 世纪 80 年代, 如文献 [17–21] 等. 最简单的微观结构噪声模型是由 Roll^[22] 提出的, 其形式如下:

$$\varepsilon_{t_k} = \alpha I_{b/s}(t_k), \quad (1.1)$$

其中 $I_{b/s}(t_k)$ 记为交易类型, 即若交易是买方发起的, 则其取值为 +1; 若交易是卖方发起的, 则其取值为 -1. 系数 α 可理解为买卖价差 (bid-ask spread) 影响的一半. 该模型后来被 Glosten 和 Harris^[23] 推广到包含交易量的情形, 具体模型如下:

$$\varepsilon_{t_k} = \alpha I_{b/s}(t_k) + \beta I_{b/s}(t_k) V_{t_k}, \quad (1.2)$$

其中 V_{t_k} 记为在时刻 t_k 的交易量. 在 (1.2) 模型的基础上, Almgren 和 Chriss^[24] 考虑了交易类型和交易率的影响, 提出了如下模型:

$$\varepsilon_{t_k} = \alpha I_{b/s}(t_k) + \beta I_{b/s}(t_k) \frac{V_{t_k}}{\Delta t_k}, \quad (1.3)$$

其中 $\Delta t_k = t_k - t_{k-1}$ 记为两个连续不断交易的持续期, $V_{t_k}/\Delta t_k$ 为交易率.

Li 等^[25] 首次在带有市场信息的微观结构噪声下, 提出了相合的积分波动率估计 ERV (estimated-price realized volatility). 但是, 他们考虑的是单个资产的积分波动率情况, 而本文研究的是多资产价格在观察不同步时的积分波动率矩阵估计问题. 在一般的扩散过程框架下, 本文无需对微观结构强加 Gauss 或次 Gauss 分布假设, 仅利用带有市场交易信息的微观结构, 提出了积分波动率矩阵估计 AERVM (averaging estimated-price realized volatility matrix), 建立了估计的大样本性质, 证明其逐点的收敛速度达到 $n^{\frac{1}{3}}$, 与 Wang 和 Zou^[14] 在无微观结构噪声的情形下 (即资产真实价格可以准确观察到) 最好估计的逐点收敛速度一致. 换言之, 本文能够利用较多的市场信息, 通过市场信息刻画波动率矩阵估计, 提高积分波动率矩阵估计的精度和效率. 另外, 在积分波动率矩阵具有稀疏性的假设下, 本文所提出的估计具有相合性. 由于本文所提出方法及其大样本理论是在一般的框架和假设条件下提出的, 所以更具一般性及更大应用范围.

本文第 2 节具体介绍所研究的观察数据模型和价格模型; 第 3 节提出 AERVM 估计方法, 及其非同步性处理技巧和估计量的调整; 第 4 节通过定理给出估计量的相合性及其收敛速度; 第 5 节通过模拟分析对 AERVM 与已存在的高维波动率矩阵估计量进行比较; 第 6 节将所提出的高维波动率矩阵

估计量应用于实际 20 只高频上海证券指数数据; 第 7 节是本文的结论, 并提出一些可能的研究方向; 第 8 节给出定理的证明.

2 模型假设

2.1 观察数据及市场信息的微观结构噪声

考虑 p 个资产, 设 $X_i(t)$ 是第 i ($i = 1, \dots, p$) 个资产在时刻 t 的真实对数价格. 在真实市场上通常很难准确地观察到 $X_i(t)$, 仅能观察到受市场微观结构噪声污染的高频数据 $Y_i(t)$, 如文献 [2, 6, 7, 12, 14, 15] 等, 它们假设第 i 个资产在时刻 $t_{i,l}$ ($l = 1, \dots, n_i$) 观察到的对数价格 $Y_i(t_{i,l})$ 满足如下加性噪声模型:

$$Y_i(t_{i,l}) = X_i(t_{i,l}) + \varepsilon_i(t_{i,l}), \quad i = 1, \dots, p, \quad l = 1, \dots, n_i, \quad (2.1)$$

其中 $X_i(t_{i,l})$ 是真实对数价格过程. 在波动率估计文献中, 通常假设市场微观结构噪声 $\varepsilon_i(t_{i,l})$ 是 Gauss 白噪声, 且独立于 $X_i(t_{i,l})$. 也有部分学者研究 $\varepsilon_i(t_{i,l})$ 的结构, 例如, Hansen 和 Lunde^[26] 考虑单变量的微观结构噪声具有序列相关性, 且该噪声结构与真实对数价格过程不是独立的; Aït-Sahalia 等^[11] 研究了相依的微观结构噪声超高频数据单资产的波动率估计, 但是他们都没考虑除资产价格外任何的市场微观结构信息.

众所周知, 高频数据包含了丰富的市场信息. 例如, 在交易数据中, 除了资产交易价格, 交易量也会被记录下来; 此外报价数据也是非常有用的公开信息; 对于个体户或公司, 他们也有额外的自身交易信息. 这些信息表明, 我们对市场微观噪声的研究不应该停留在简单的加性微观噪声模型 (2.1) 中, 而应当考察较为复杂广泛的市场微观噪声模型. 因此, 本文考虑如下模型:

$$Y_i(t_{i,l}) = X_i(t_{i,l}) + g(\mathbf{Z}_i(t_{i,l}); \boldsymbol{\theta}_i), \quad i = 1, \dots, p, \quad l = 1, \dots, n_i, \quad (2.2)$$

其中 $\mathbf{Z}_i(t_{i,l})$ 是市场信息集, 其包含交易量、交易类型、交易率和买卖价差等; $\boldsymbol{\theta}_i$ 是一个有限维参数; $g(\cdot)$ 表示的是市场微观结构噪声, 该函数可以是关于参数 $\boldsymbol{\theta}_i$ 的线性函数也可以是非线性函数. 它是将市场信息集 $\mathbf{Z}_i(t_{i,l})$ 作为协变量的已知参数函数, 引言中的模型 (1.1)–(1.3) 是 $g(\mathbf{Z}_i(t_{i,l}); \boldsymbol{\theta}_i)$ ($i = 1, \dots, p$, $l = 1, \dots, n_i$) 的特例形式. 通过函数 g 可以充分利用高频数据信息帮助我们提高积分波动率矩阵估计的精度. 此模型结构最早是由 Li 等^[25] 在估计单资产积分波动率矩阵时提出来的.

2.2 价格模型

设 $\mathbf{X}(t) = (X_1(t), \dots, X_p(t))^T$ 是 p 个资产在时间 t 的真实对数价格, 采用文献 [2, 11, 14, 15, 25] 中的价格模型, 假设 $\mathbf{X}(t)$ 服从连续时间的扩散过程:

$$d\mathbf{X}(t) = \boldsymbol{\mu}_t dt + \boldsymbol{\sigma}_t^T d\mathbf{B}_t, \quad t \in [0, T], \quad (2.3)$$

其中 $\boldsymbol{\mu}_t = (\mu_1(t), \dots, \mu_p(t))^T$ 是漂移向量过程, $\mathbf{B}_t = (B_{1t}, \dots, B_{pt})^T$ 是 p 维 Brown 运动 (即 B_{it} ($i = 1, \dots, p$) 是相互独立的 Brown 运动) 且 $\boldsymbol{\sigma}_t$ 是 $p \times p$ 的随机矩阵过程. $[0, T]$ 为观察时间, 不失一般性, 假设 $T = 1$, $\mathbf{X}(t)$ 的点波动率矩阵是

$$\boldsymbol{\gamma}(t) = (\gamma_{ij}(t))_{1 \leq i, j \leq p} = \boldsymbol{\sigma}_t^T \boldsymbol{\sigma}_t. \quad (2.4)$$

其二次变差和该二次变差的极限分别为

$$[\mathbf{X}, \mathbf{X}]_t = \sum_{0 \leq t_l \leq t_{l+1} \leq t} (\mathbf{X}(t_{l+1}) - \mathbf{X}(t_l))(\mathbf{X}(t_{l+1}) - \mathbf{X}(t_l))^T,$$

$$\lim_{\Delta \rightarrow 0} [\mathbf{X}, \mathbf{X}]_t = \int_0^t \boldsymbol{\gamma}(s) ds = \left(\int_0^t \gamma_{ij}(s) ds \right)_{1 \leq i, j \leq p}, \quad t \in [0, 1],$$

其中 $\Delta = \max\{t_{l+1} - t_l, l = 1, 2, \dots, n\}$. 因此, $\mathbf{X}(t)$ 的二次变差极限是积分波动率矩阵. 由于金融资产的点波动率矩阵具有较强的记忆性和时变性, 很难精准地获得, 所以通常利用资产的积分波动率矩阵估计代替对资产点波动率矩阵的准确估计, 这种方法是制定更为合理投资决策的一种重要手段. 于是估计积分波动率矩阵是我们的一个重要目标, 即估计

$$\boldsymbol{\Gamma} = (\Gamma_{ij})_{1 \leq i, j \leq p} = \int_0^1 \boldsymbol{\gamma}(s) ds = \left(\int_0^1 \gamma_{ij}(s) ds \right)_{1 \leq i, j \leq p}.$$

通常估计积分波动率矩阵将会面临三大困难: (1) 高频观察数据的非同步性. 在实际中, 不同资产的交易时间和交易量都是不一样的, 而我们估计的积分波动率矩阵反映的是同一时间或同一时间段各个资产的联合波动情况. (2) 高维矩阵估计的相合性. 高维矩阵估计的 L_∞ -范数的相合性是比较容易建立的, 但是 L_2 -范数的相合性是非常困难的, 往往需要假设高维矩阵本身具有某种稀疏性. 文献 [14–16] 强加某些稀疏性技巧保证估计量是 L_2 相合的. (3) 保证估计矩阵的半正定性. 点波动率矩阵事实上就是点协方差矩阵, 本身具有半正定性, 因此, 积分波动率矩阵也具有半正定性. 文献 [14–16] 虽然通过假设稀疏性保证估计的相合性, 但是它们并没有保证该相合估计量的半正定性, 这与积分波动率的性质是相悖的. 本文将克服上述三大困难, 在带有市场信息的微观结构噪声下对积分波动率矩阵进行估计.

3 积分波动率矩阵的估计

3.1 参数估计

在无市场微观噪声的情形下, 已实现波动率 (RV, realized volatility)

$$RV := \sum_{k=1}^n (\Delta X_{t_k})^2 \quad (3.1)$$

通常是单资产积分波动率估计的一个最简单而有效的估计量, 最早是由 Andersen 等^[1] 提出来的, 该估计利用高频日内数据进行直接估计, 没有假设参数模型, 克服了当资产数很多时传统的多元 GARCH 的“维数灾难”. 在模型 (2.2) 和 (2.3) 下, 类似于文献 [2, 3, 14, 16] 中的 MSRV 等对积分波动率的估计, 本文将基于已实现波动率估计方法对积分波动率矩阵进行估计.

首先根据模型 (2.2) 和 (2.3), 对参数 $\boldsymbol{\theta}_i$ ($i = 1, \dots, p$) 进行估计. 记 $\Delta X_i(t_{i,l}) = X_i(t_{i,l}) - X_i(t_{i,l-1})$, $\Delta Y_i(t_{i,l}) = Y_i(t_{i,l}) - Y_i(t_{i,l-1})$ 和 $\Delta g(\mathbf{Z}_i(t_k); \boldsymbol{\theta}_i) = g(\mathbf{Z}_i(t_k); \boldsymbol{\theta}_i) - g(\mathbf{Z}_i(t_{k-1}); \boldsymbol{\theta}_i)$, 则有

$$\Delta Y_i(t_{i,l}) = \Delta g(\mathbf{Z}_i(t_{i,l}); \boldsymbol{\theta}_i) + \Delta X_i(t_{i,l}). \quad (3.2)$$

为估计参数 $\boldsymbol{\theta}_i$, 可以利用的资产价格模型特征, 将模型 (3.2) 视作非线性回归模型, 即在真实对数价格在相对小的时间内变动很小时, $\Delta X_i(t_{i,l})$ 作为“噪声”, $\Delta g(\mathbf{Z}_i(t_{i,l}); \boldsymbol{\theta}_i)$ 作为“信号”, 利用模型 (3.2)

对参数 θ_i 进行估计. 考虑简单的情形, 即模型 (2.3) 中漂移项 $\mu_i(t) = 0$, 波动项 $\sigma_{ii}(t) = \sigma_i$, 观察的时间点是均匀的, 即 $t_{i,l} = j/n_i$ ($l = 0, 1, \dots, n_i, i = 1, \dots, p$), “噪声”项 $\Delta X_i(t_{i,l})$ 是均值为 0 的误差, 在模型 (2.3) 下其近似服从 $N(0, \sigma_i^2/n_i)$. 因此, 如果 $\Delta X_i(t_{i,l})$ 独立于 $\Delta g(\mathbf{Z}_i(t_{i,l}); \theta_i)$, θ_i 的极大似然估计 $\hat{\theta}_i$ 为

$$\begin{aligned}\hat{\theta}_i &= \arg \min_{\theta_i} Q_{n_i}(Y_i, \mathbf{Z}_i, \theta_i), \quad i = 1, \dots, p, \\ Q_{n_i}(Y_i, \mathbf{Z}_i, \theta_i) &= \frac{1}{2} \sum_{l=1}^{n_i} (\Delta Y_i(t_{i,l}) - \Delta g(\mathbf{Z}_i(t_{i,l}); \theta_i))^2.\end{aligned}$$

(1.1)–(1.3) 是 g 的线性表达形式 (即 $g(\mathbf{Z}_i(t_{i,l}); \theta_i) = \mathbf{Z}_i^T(t_{i,l})\theta_i$). 此时, $\hat{\theta}_i$ 的精确表达式为

$$\hat{\theta}_i^L = \left(\sum_{l=1}^{n_i} \Delta \mathbf{Z}_i^T(t_{i,l}) \Delta \mathbf{Z}_i(t_{i,l}) \right)^{-1} \left(\sum_{l=1}^{n_i} \Delta \mathbf{Z}_i^T(t_{i,l}) \Delta Y_i(t_{i,l}) \right), \quad i = 1, \dots, p.$$

当 g 是一般的非线性函数时, 可以使用非线性最小二乘法获得参数 θ 的估计. 为了克服计算困难, 通常先对 g 的结构形式进行逼近, 然后再利用最小二乘法估计参数 θ .

在一定的假设条件下 (见假设 4.1), 极大似然估计量 $\hat{\theta}_i$ 收敛到 θ_i , 且 $\hat{\theta}_i - \theta_i = O_p(1/n)$, 具体可参见文献 [25, 定理 1]. 从而得到估计

$$\Delta \hat{X}_i(t_{i,l}) = \Delta Y_i(t_{i,l}) - \Delta g(\mathbf{Z}_i(t_{i,l}); \hat{\theta}_i).$$

根据已实现波动率的定义 (3.1), $\Delta X_i(t_{i,l})$ 的估计量为 $\Delta \hat{X}_i(t_{i,l})$, 基于已实现波动率估计方法将容易获得积分波动率矩阵估计.

3.2 已实现波动率矩阵

现实生活中的高频数据, 通常存在非同步性 (不同资产交易发生的时间不同, 典型地 $t_{i,l} \neq t_{j,l}$, 对某些的 $i \neq j$) 的问题, 我们采用 Zhang^[10] 的 PT (previous-tick) 方法, 首先对于固定的整数 m , 取 τ_r ($r = 1, \dots, m$) 作为预先取定的样本频率, 然后记 $\boldsymbol{\tau} = \{\tau_r, r = 1, \dots, m\}$. 对于资产 i , 定义 PT 时间为

$$\tau_{i,r} = \max\{t_{i,l} \leq \tau_r, l = 1, \dots, n_i\}, \quad r = 1, \dots, m.$$

也就是说, 对于任意固定的 τ_r ($r = 1, \dots, m$), 我们取不大于 τ_r 且离 τ_r 最近一次资产 i 发生的交易时间 $\tau_{i,r}$ 看作资产 i 在 τ_r 发生交易. 这些 $\tau_{i,r}$ ($i = 1, \dots, p$) 虽然并不相同, 但是它们都非常接近 τ_r , 所以, 我们认为这些交易是同时发生的, 从而达到了交易时间的同步化目的.

基于 $\boldsymbol{\tau}$ 可以定义资产 i 和 j 的已实现协波动率为

$$\begin{aligned}\hat{\Gamma}_{ij}(\boldsymbol{\tau}) &= \sum_{r=1}^m [\Delta \hat{X}_i(\tau_{i,r})][\Delta \hat{X}_j(\tau_{j,r})] \\ &= \sum_{r=1}^m [\Delta Y_i(\tau_{i,r}) - \Delta g(\mathbf{Z}_i(\tau_{i,r}); \hat{\theta}_i)][\Delta Y_j(\tau_{j,r}) - \Delta g(\mathbf{Z}_j(\tau_{j,r}); \hat{\theta}_j)],\end{aligned}\tag{3.3}$$

则已实现波动率矩阵为

$$\hat{\Gamma}(\boldsymbol{\tau}) = (\hat{\Gamma}_{ij}(\boldsymbol{\tau}))_{i,j=1,\dots,p}.\tag{3.4}$$

预先选定样本频率 τ (通常选用等距的网格点形式), 对于固定的 m , 存在 $K = [n/m]$ 类不重叠的等距网格,

$$\tau^k = \left\{ \frac{r}{m}, r = 1, \dots, m \right\} + \frac{k-1}{n} = \left\{ \frac{r}{m} + \frac{k-1}{n}, r = 1, \dots, m \right\},$$

其中 $k = 1, \dots, K$, n 是平均样本, 即

$$n = \frac{1}{p} \sum_{i=1}^p n_i.$$

对每个样本频率 τ^k , 使用 (3.3) 和 (3.4) 定义的资产 i 和 j 的已实现协波动率 $\hat{\Gamma}_{ij}(\tau^k)$ 和已实现波动率矩阵 $\hat{\Gamma}(\tau^k)$. 进而定义平均已实现协波动率 $\hat{\Gamma}_{ij}$ 和平均已实现波动率矩阵 $\hat{\Gamma}$, 即

$$\hat{\Gamma}_{ij} = \frac{1}{K} \sum_{k=1}^K \hat{\Gamma}_{ij}(\tau^k), \quad \hat{\Gamma} = (\hat{\Gamma}_{ij}) = \frac{1}{K} \sum_{k=1}^K \hat{\Gamma}(\tau^k). \quad (3.5)$$

其估计形式和 TSRV^[2] 单变量的情形类似, $\hat{\Gamma}$ 是 K 个已实现波动率矩阵 $\hat{\Gamma}(\tau^k)$ 的平均, 即首先利用 τ^k 的子样数据计算 $\hat{\Gamma}(\tau^k)$, 然后用它们的平均定义 $\hat{\Gamma}$, 此估计量称为 AERVM (averaging estimated-price realized volatility matrix), 这种子样平均方法的目的是为了减少市场微观结构噪声的影响, 进而获得更好的积分波动率矩阵估计.

高频金融数据通常既不是等空间的也不是同步的, 因此, 一些资产在某些时间区间上观察的密度会多于或少于其他资产. 也就是说, 一个资产在两个连续样本频率的时间点上没有观察值, 那么 (3.3) 涉及这两个连续时间点时, $\Delta \hat{X}_i(\tau_{i,l})$ 就为零. 从 (3.5) 可以发现我们提出的积分波动率矩阵估计量 AERVM 对不同观察密度的数据自动进行了调整.

3.3 AERVM 的调整

当 p 比较小时, AERVM $\hat{\Gamma}$ 是 Γ 的一个很好的估计量; 当 p 比较大时, Wang 和 Zou^[14] 举例说明了在无噪声的情形下, 对于 (2.4) 中的 $\gamma(t)$, 当 n 和 p 趋向于无穷时, 基于已实现波动率估计方法得到的未调整的估计量通常都是不相合的. 特别地, 当 p 非常大时, 它们的特征值和特征向量与真实 Γ 的特征值和特征向量相差甚远. 因此, 需要对 AERVM $\hat{\Gamma}$ 进行适当的调整, 从而使 $\hat{\Gamma}$ 是 Γ 的相合估计. 根据 Bickel 和 Levina^[27, 28] 的思想, 本文考虑 Γ 在退化条件假设或稀疏条件假设下积分波动率矩阵估计, 即假设真实的积分波动率矩阵 Γ 具有带状特征或门限特征.

退化条件: 除去 Γ 对角线上的元素, 假设其非对角线元素满足

$$|\Gamma_{ij}| \leq \frac{M}{1 + |i - j|^{\alpha+1}}, \quad 1 \leq i, j \leq p, \quad E[M] \leq C, \quad (3.6)$$

其中 M 为正随机变量, C 和 α 为一般的正常数.

稀疏条件: 假设 Γ 满足

$$\sum_{j=1}^p |\Gamma_{ij}|^\delta \leq M\pi(p), \quad i = 1, \dots, p, \quad E[M] \leq C, \quad (3.7)$$

其中 M 为正随机变量, $0 \leq \delta < 1$; $\pi(p)$ 是 p 的慢变函数, 如 1 、 $\ln(p)$ 和 $\ln \ln(p)$ 等.

注 3.1 当 $\delta = 0$ 时, $\mathbf{\Gamma}$ 的每一行元素的和最多为 $M\pi(p)$. 事实上, 当 $\delta = 1/(\alpha + 1)$, $\pi(p) = \ln(p)$ 或 $1/(\alpha + 1) \leq \delta < 1$, $\pi(p) = 1$ 时, 退化条件是稀疏条件的特殊情形.

注 3.2 在真实市场中, 通过对数价格向量 $\mathbf{Y}(t)$ 建立的积分波动率矩阵估计量退化条件是不可能满足的. 因为退化条件依赖于 p 个资产的排列顺序, 而当股票没经过自然排序时, 随着排列顺序的不同, 所得到股票收益的积分波动率矩阵估计量也会有很大差异. 波动率矩阵估计在金融应用中, 稀疏条件假设比退化条件假设更实际, 因为稀疏矩阵包括分块对角矩阵、对角退化元素矩阵、每行或每列相对较小的非零元素矩阵, 以及随机排列前面三种矩阵的行和列所得到的矩阵. 稀疏性具有一定的实际意义, 它保留了在驱动波动率运动和捕捉市场风险中起到主导性作用的相对少数的市场因子, 从而方便投资者有针对性地进行投资.

对于 $\mathbf{\Gamma}$ 满足退化条件 (3.6), 它的较大元素都落在它对角线位置上, 其他位置的元素是可以忽略的. 于是调整 $\hat{\mathbf{\Gamma}}$, 给出定义

$$\mathcal{B}_b[\hat{\mathbf{\Gamma}}] = (\hat{\Gamma}_{ij}I(|i - j| \leq b)),$$

其中 b 是宽带参数, $I(|i - j| \leq b)$ 是示性函数, 称该波动率矩阵 $\mathcal{B}_b[\hat{\mathbf{\Gamma}}]$ 为 BAERVM 估计量.

如果 $\mathbf{\Gamma}$ 满足稀疏条件 (3.7), 它的重要的元素是 $\mathbf{\Gamma}$ 中那些绝对值超过门限的元素. 给出定义

$$\mathcal{T}_\varpi[\hat{\mathbf{\Gamma}}] = (\hat{\Gamma}_{ij}I(|\hat{\Gamma}_{ij}| \geq \varpi)),$$

其中 ϖ 是门限, 称该波动率矩阵 $\mathcal{T}_\varpi[\hat{\mathbf{\Gamma}}]$ 为 TAERVM 估计量.

与多数已存在的波动率矩阵估计量一样, 虽然并不能保证有限样本的估计量 BAERVM 和估计量 TAERVM 的半正定性, 但有限样本的估计量 AERVM 的半正定性¹⁾ 是可以保证的. 当样本量和资产数趋向于无穷时, 估计量 BAERVM 和 TAERVM 是渐近半正定的.

4 估计的渐近性质

为方便叙述本文所提出估计量的渐近性质, 设 $\mathbf{x} = (x_1, \dots, x_p)^T$ 是 p 维向量, $\mathbf{U} = (U_{ij})$ 是 $p \times p$ 矩阵. 定义 l_d -范数:

$$\|\mathbf{x}\|_d = \left(\sum_{i=1}^p |x_i|^d \right)^{1/d}, \quad \|\mathbf{U}\|_d = \sup\{\|\mathbf{U}\mathbf{x}\|_d, \|\mathbf{x}\|_d = 1\}, \quad d = 1, 2, \infty.$$

矩阵模 $\|\mathbf{U}\|_2$ 是 $\mathbf{U}\mathbf{U}^T$ 的最大特征值的平方根, 且

$$\|\mathbf{U}\|_1 = \max_{1 \leq j \leq p} \sum_{i=1}^p |U_{ij}|, \quad \|\mathbf{U}\|_\infty = \max_{1 \leq i \leq p} \sum_{j=1}^p |U_{ij}|.$$

通常 $\|\mathbf{U}\|_1$ 和 $\|\mathbf{U}\|_\infty$ 分别称为列和范数与行和范数. 对于对称的 $\mathbf{U}$, 它的谱范数等于最大绝对特征值, 且 $\|\mathbf{U}\|_2 \leq \|\mathbf{U}\|_1 = \|\mathbf{U}\|_\infty$. 本文主要涉及 l_d -范数, 其中 $d = 1, 2, \infty$.

为了描述提出估计量的渐近性质, 本文做了以下假设.

假设 4.1 对于 $i = 1, \dots, p$, 有

(i) 向量 $\boldsymbol{\mu}_t$ 和矩阵 $\boldsymbol{\sigma}_t$ 中的每个元素都是局部有界的, 且 $\inf_{t \in [0, 1]} \sigma_{ii}(t) > 0$, a.s.;

1) 根据半正定性定义以及估计量 AERVM 的定义形式容易获得估计量的半正定性.

- (ii) 对所有的 $\tau_{i,r}, r = 1, \dots, m$, 在给定 $\mathcal{F}_{\tau_{i,r}}$ 下, $\mathbf{Z}_{\tau_{i,r}}$ 与 $\Delta X_{\tau_{i,r}}$ 是相互独立的;
- (iii) $\max_{i,r} |\mathbf{Z}_{\tau_{i,r}}| = O_p(1)$, 其中 $|\mathbf{Z}|$ 表示向量 $\mathbf{Z}$ 的欧氏距离;
- (iv) 对于 $q \in \mathbb{N}$, $\boldsymbol{\theta}_i$ 的参数空间 Θ 在 $\mathbb{R}^q$ 上是紧集, 且当 $\boldsymbol{\theta}_i$ 在邻域 $\mathcal{N}(\boldsymbol{\theta}_{i0}) \subset \Theta$ 内时, $g(\mathbf{z}_i; \boldsymbol{\theta}_i)$ 关于 $\boldsymbol{\theta}_i$ 是二阶连续可微的;
- (v) 对所有的 $\boldsymbol{\theta}_{i1}, \boldsymbol{\theta}_{i2} \in \mathcal{N}(\boldsymbol{\theta}_{i0})$, $|g(\mathbf{z}_i; \boldsymbol{\theta}_{i1}) - g(\mathbf{z}_i; \boldsymbol{\theta}_{i2})| \leq L_0(\mathbf{z}_i)|\boldsymbol{\theta}_{i1} - \boldsymbol{\theta}_{i2}|$, 其中 $L_0(\mathbf{z}_i)$ 是局部有界的;
- (vi) 对所有的 $\boldsymbol{\theta}_i \in \mathcal{N}(\boldsymbol{\theta}_{i0})$, $|\frac{\partial g}{\partial \boldsymbol{\theta}_i}(\mathbf{z}_i; \boldsymbol{\theta}_i)| \leq f(\mathbf{z}_i)$, 其中 $f(\mathbf{z}_i)$ 是局部有界的;
- (vii) $|\frac{\partial g}{\partial \boldsymbol{\theta}_i}(\mathbf{z}_i; \boldsymbol{\theta}_{i1}) - \frac{\partial g}{\partial \boldsymbol{\theta}_i}(\mathbf{z}_i; \boldsymbol{\theta}_{i2})| \leq L_1(\mathbf{z}_i)|\boldsymbol{\theta}_{i1} - \boldsymbol{\theta}_{i2}|$, 其中 $L_1(\mathbf{z}_i)$ 是局部有界的;
- (viii) 对 $\forall \varepsilon > 0$, 当 $m \rightarrow \infty$ 时,

$$\inf_{|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i0}| \geq \varepsilon} \sum_{r=1}^m |\Delta g(\mathbf{Z}_i(\tau_{i,r}); \boldsymbol{\theta}_i) - \Delta g(\mathbf{Z}_i(\tau_{i,r}); \boldsymbol{\theta}_{i0})|^2 \rightarrow \infty, \quad \text{a.s.};$$

(ix)

$$\left\| \left[\frac{1}{m} \sum_{r=1}^m \Delta \frac{\partial g}{\partial \boldsymbol{\theta}_i}(\mathbf{Z}_i(\tau_{i,r}); \boldsymbol{\theta}_{i0}) \Delta \frac{\partial g}{\partial \boldsymbol{\theta}_i}(\mathbf{Z}_i(\tau_{i,r}); \boldsymbol{\theta}_{i0}) \right]^{-1} \right\|_2 = O_p(1),$$

其中对任意的矩阵 A , $\|A\|_2$ 表示谱范数或 l_2 -范数;

(x) 对任意的 $k, l = 1, \dots, q$, 且 $\boldsymbol{\theta}_{i1}, \boldsymbol{\theta}_{i2} \in \mathcal{N}(\boldsymbol{\theta}_{i0})$, 有

$$\left| \frac{\partial^2 g}{\partial \theta_{ik} \partial \theta_{il}}(\mathbf{z}_i; \boldsymbol{\theta}_{i1}) - \frac{\partial^2 g}{\partial \theta_{ik} \partial \theta_{il}}(\mathbf{z}_i; \boldsymbol{\theta}_{i2}) \right| \leq L_2(\mathbf{z}_i)|\boldsymbol{\theta}_{i1} - \boldsymbol{\theta}_{i2}|,$$

其中 $L_2(\mathbf{z}_i)$ 是局部有界的;

(xi) 对任意的 $k, l = 1, \dots, q$, $\boldsymbol{\theta}_i \in \mathcal{N}(\boldsymbol{\theta}_{i0})$, 有

$$\left| \frac{\partial^2 g}{\partial \theta_{ik} \partial \theta_{il}}(\mathbf{z}_i; \boldsymbol{\theta}_i) \right| \leq L_3(\mathbf{z}_i),$$

其中 $L_3(\mathbf{z}_i)$ 是局部有界的.

注 4.1 假设 (ii) 类似于通常的回归假设, 在模型 (2.2) 下, 相当于假设资产的下一次瞬时交易仅依赖最新的交易市场信息. 当高频交易发生时, 尤其在算法交易的过程中, 资产的买卖与否以及交易量完全由过去交易决定, 因此, 该假设是合理的. 在极大似然估计 (MLE) 文献中, 假设 (iv)–(xi) 是很常规的; 假设 (viii) 类似于 MLE 的可识别条件; 假设 (ix) 与 Fisher 信息矩阵的可逆条件一致. 以上的假设主要是为了保证极大似然估计量 $\hat{\boldsymbol{\theta}}_i$ 收敛到 $\boldsymbol{\theta}_i$, 具体参见文献 [25, 定理 1].

假设 4.2 对于 $\beta \geq 2$, 模型 (2.3) 中的漂移项和 $\mathbf{X}(t)$ 的点波动率矩阵 (2.4) 的对角元素分别满足

$$\max_{1 \leq i \leq p} \max_{0 \leq t \leq 1} \mathbb{E}[|\mu_i(t)|^{2\beta}] < \infty, \quad \max_{1 \leq i \leq p} \max_{0 \leq t \leq 1} \mathbb{E}[|\gamma_{ii}(t)|^\beta] < \infty.$$

假设 4.3 存在常数 C_1 和 C_2 使得

$$\max_{1 \leq i \leq p} \frac{n_i}{n} \leq C_1, \quad \max_{1 \leq i \leq p} \max_{1 \leq l \leq n_i} |t_{i,l} - t_{i,l-1}| \leq \frac{C_2}{n},$$

其中 $n = (n_1 + \dots + n_p)/p$, n_i 是第 i 个资产的观察样本量, $t_{i,l}$ 是第 i 个资产第 l 个观察样本的记录时间.

注 4.2 假设 4.2 强调 $\mathbf{X}(t)$ 的漂移项和扩散协方差的 2β 阶矩存在性, 是由于金融数据总是具有有限阶矩, 这个假设是符合实际的. 假设 4.3 的第一个不等式是对 p 个资产数据样本频率进行限制, 目的是为保证在整个观察区间内没有“洞”. 事实上, 如果存在“洞”, 那么没有方法能够保证波动率矩阵估计在“洞”上是相合的. 假设 4.3 的第二个不等式是为了保证在连续的两个预先确定的样本区间里有足够多的不同观察值对积分波动率矩阵进行估计, 参见文献 [16].

考虑市场微观结构噪声是市场信息的参数函数, 利用高频金融数据所提出的高维波动率矩阵估计量具有如下渐近性质.

定理 4.1 在模型 (2.2)、(2.3) 和假设 4.1–4.3 下, 对所有的 $1 \leq i, j \leq p$, 高维波动率矩阵估计量 AERVM 满足

$$E|\hat{\Gamma}_{ij} - \Gamma_{ij}|^\beta \leq Cn^{-\frac{\beta}{3}}, \quad (4.1)$$

其中 C 是与 n 和 p 无关的常数. 这里的 $\beta \geq 2$, 但是当 $\beta = 1$ 时,

$$E|\hat{\Gamma}_{ij} - \Gamma_{ij}| \leq \sqrt{E|\hat{\Gamma}_{ij} - \Gamma_{ij}|^2} \leq Cn^{-\frac{1}{3}},$$

(4.1) 也成立.

注 4.3 定理 4.1 表明 AERVM 估计量的每个元素收敛速度都是 $n^{-1/3}$, 相当于文献 [14] 中无噪声的情形. 这说明运用更多信息能够提高估计积分波动率矩阵的准确度.

定理 4.2 假设 $\hat{\Gamma}$ 满足退化条件 (3.6), 在模型 (2.2)、(2.3) 和假设 4.1–4.3 下, 有

$$\|\mathcal{B}_b[\hat{\Gamma}] - \Gamma\|_2 \leq \|\mathcal{B}_b[\hat{\Gamma}] - \Gamma\|_\infty = O_P((n^{-\beta/3}p)^{\alpha/(\alpha\beta+\beta+1)}),$$

其中宽带参数 b 的阶数为 $(n^{-\beta/3}p)^{-1/(\alpha\beta+\beta+1)}$.

定理 4.3 假设 $\hat{\Gamma}$ 满足稀疏条件 (3.7), 在模型 (2.2)、(2.3) 和假设 4.1–4.3 下, 有

$$\|\mathcal{T}_\varpi[\hat{\Gamma}] - \Gamma\|_2 \leq \|\mathcal{T}_\varpi[\hat{\Gamma}] - \Gamma\|_\infty = O_P(\pi(p)\varpi^{1-\delta}),$$

其中门限参数 $\varpi = n^{-1/3}p^{2/\beta}h_{n,p}$, $h_{n,p}$ 是任意一个以很慢的速度收敛到无穷的序列, 例如,

$$h_{n,p} = \ln \ln(n \wedge p).$$

注 4.4 当 Γ 满足退化条件时, 也就是相当于 $\delta = 1/(1+\alpha)$ 和 $\pi(p) = \ln(p)$ 时的稀疏条件, 由定理 4.3 可以知道主导元素的阶为 $(n^{-1/3}p^{2/\beta})^{\alpha/(\alpha+1)}$, 与定理 4.2 比较, 我们会发现选取合适的 β 两收敛速度是非常接近的.

注 4.5 因为 $h_{n,p}$ 和 $\pi(p)$ 皆为慢变函数, 当 p 增长速度慢于 $n^{\beta/6}$ 时, 定理 4.3 中的式子趋向于零的收敛速度近似为 $(n^{-1/3}p^{2/\beta})^{1-\delta}$. 此外, Wang 和 Zou [14] 及 Tao 等 [16] 也提出了波动率矩阵的估计量分别为门限 ARVM 和门限 MSRVM, 然而他们所提出的波动率矩阵估计量的收敛速度分别为 $\pi(p)(n^{-1/6}p^{2/\beta})^{1-\delta}$ 和 $\pi(p)(n^{-1/4}p^{2/\beta})^{1-\delta}$, 皆慢于本文在定理 4.3 中提出的波动率矩阵的门限 AERVM 的收敛速度.

5 模拟研究

5.1 模拟模型

本节通过数值模拟检验本文所提出的积分波动率矩阵估计量的统计性质. 考虑下面三种形式的微观结构噪声模型:

$$\begin{aligned}
g_1(V_{t_k}, I_{b/s}(t_k); \alpha, \beta) &= \alpha I_{b/s}(t_k) + \beta I_{b/s}(t_k) \frac{V_{t_k}}{\Delta t_k}, \\
g_2(V_{t_k}, I_{b/s}(t_k); \alpha, \beta^+, \beta^-) &= \begin{cases} \alpha I_{b/s}(t_k) + \beta^+ I_{b/s}(t_k) \frac{V_{t_k}}{\Delta t_k}, & \text{买家发起,} \\ -\alpha I_{b/s}(t_k) - \beta^- I_{b/s}(t_k) \frac{V_{t_k}}{\Delta t_k}, & \text{卖家发起,} \end{cases} \\
g_3(V_{t_k}, I_{b/s}(t_k); \beta, \gamma) &= I_{b/s}(t_k) \ln \left(\gamma + \beta I_{b/s}(t_k) \frac{V_{t_k}}{\Delta t_k} \right),
\end{aligned}$$

其中 $I_{b/s}(t_k)$ 记为交易类型, 即若交易是买方发起的, 则其取值为 +1; 若交易是卖方发起的, 则其取值为 -1. V_{t_k} 记为在时刻 t_k 的交易量, $\Delta t_k = t_k - t_{k-1}$ 记为两个连续不断交易的持续期, $V_{t_k}/\Delta t_k$ 为交易率.

在模拟当中, 假定 $\alpha = 6.25 \times 10^{-4}$, $\beta = \beta^+ = 2.50 \times 10^{-12}$, $\beta^- = 1.00 \times 10^{-12}$ 和 $\gamma = 1 + \alpha = 1.000625$, 其中模型 g_1 同 Almgren 和 Chriss^[24] 所考虑的买卖发起方和交易率的线性影响, 即 (1.2); 模型 g_2 考虑了交易率和非对称的买卖发起方的影响; 模型 g_3 则是考虑交易率的非线性影响. Keim 和 Madhavan^[29] 发现对于买方而言噪声模型关于交易率是凹的, 而对于卖家而言噪声模型关于交易率是凸的. 事实上, 当交易率非常低时, 模型 g_3 近似于模型 g_1 ; 然而当交易率非常高时, 模型 g_3 增长将慢于线性模型 g_1 的增长.

价格模型设定为漂移项为零的模型 (2.3), 即

$$d\mathbf{X}(t) = \boldsymbol{\sigma}_t^T d\mathbf{B}_t, \quad t \in [0, 1],$$

其中 $\mathbf{B}_t = (B_{1t}, \dots, B_{pt})$ 是标准 p 维 Brown 运动; $\boldsymbol{\gamma}(t) = (\gamma_{ij}(t))_{1 \leq i, j \leq p} = \boldsymbol{\sigma}_t^T \boldsymbol{\sigma}_t$, 设 γ_t 的对角元素服从 $(0, 0.1)$ 上的均匀分布, 在给定 γ_t 对角元素, 定义非对角元素为

$$\gamma_{i,j}(t) = \kappa(t)^{|i-j|} \sqrt{\gamma_{i,i}(t)\gamma_{j,j}(t)},$$

其中过程 $\kappa(t) = (e^{2u(t)} - 1)/(e^{2u(t)} + 1)$, $du(t) = 0.03[0.64 - u(t)]dt + 0.118u(t)dW_{\kappa,t}$, $W_{\kappa,t} = \sqrt{0.96}W_{\kappa,t}^0 - 0.2 \sum_{i=1}^p B_{it}/\sqrt{p}$; $W_{\kappa,t}^0$ 是独立于 $\mathbf{B}_t$ 的一维标准 Brown 运动, 这些参数的设定类似于文献 [14, 16].

交易类型 $I_{b/s}$ 是由概率为 $\frac{1}{2}$ 的 Bernoulli 过程生成的, 我们认为买卖双方是以相同的概率进行交易的. 对于交易量的模型可以参见文献 [25], 建模如下:

$$V_{t_k} = H_{t_k} + e_{t_k},$$

其中 H_{t_k} ($k = 0, 1, \dots, n$) 是独立的随机变量, 服从均值为 $\int_{t_{k-1}}^{t_k} h_t dt$ 、标准差为 5000 的伽马分布, $h_t = (0.025(t - 0.5)^2 + 0.05) \times 10^8$; $\{e_{t_k}\}$ 服从 ARMA(1, 1) 模型

$$e_{t_k} = \phi_1 e_{t_{k-1}} + \psi_1 u_{t_{k-1}},$$

其中 $\phi_1 = 0.5$, $\psi_1 = 0.5$, $\{u_{t_k}\}$ 是独立同分布的, 服从均值为零、标准差为 100 的正态分布.

首先使用模型生成数据, 然后利用文献 [25] 中给出的方法估计噪声模型中的参数, 从而获得估计噪声; 其次利用 PT 技巧处理多资产的非同步性计算估计量 AERVM 和门限估计量 TAERVM, 并计算 Wang 和 Zou^[14] 的估计量 ARVM 和 TARVM 以及 Tao 等^[16] 的估计量 MSRV 和 TMSRV. 由于退化条件是稀疏条件的特例, 并且在现实中研究不同资产的积分波动率矩阵时资产具有无序性, 所以退化条件是不符合实际的. 故而在模拟实验中只研究了稀疏条件的积分波动率矩阵估计量.

5.2 模拟结果

在模拟研究中, 将使用本文提出的方法估计积分波动率矩阵, 取 $n = 5000$, $p = 3$ 或 $p = 50$, 整个模拟过程重复 200 次. 积分波动率矩阵估计量的均方差 (MSE) 指的是该估计量与真实积分波动率矩阵作差的 l_2 -范数. 为说明本文所给方法的优越性, 本文将所提出的估计量与 Wang 和 Zou^[14] 的积分波动率估计量 ARVM 和 TARVM 以及 Tao 等^[16] 的积分波动率估计量 MSRVM 和 TMSRVM 的均方差 (MSE, mean square error) 和标准差 (se, standard error) 进行比较. 一般来讲, 观察到的收益受市场微观结构噪声的影响, 特别是市场信息, 所以, 本文的模拟是在包含交易信息的不同的市场微观结构噪声模型下进行估计量比较的. 主要的模型参数设置参见文献 [14, 25], 为方便观察, 本文对得到的估计量扩大一万倍后进行计算得到均方差和标准差. 如无特殊说明, 本文所呈现的数据和图形数据都是将估计量扩大了一万倍后计算得到.

表 1-3 给出了 $p = 3$ 个资产在不同市场微观结构噪声模型和不同的噪声水平下 6 种积分波动率估计量 (即 ARVM 和 TARVM, MSRVM 和 TMSRVM, AERVM 和 TAERVM) 的均方差 MSE 和标准差 se. 本文中表格里的数值均表示估计量的均方差, 括号里的数值均表示估计量的标准差. 为了方便比较这 6 种积分波动率估计量, 表 1-3 给出的都是在选取最优的 K 、 M 、 N 和 ϖ 情形下得到 6 种最优的估计量进行比较的. 同时, 经过多次模拟验证对于同一个真实的波动率矩阵, 本文所提出来的积分波动率矩阵估计量 AERVM 和 TAERVM 的均方差和标准差在不同的噪声水平下是不发生改变的, 然而估计量 ARVM 和 TARVM 以及估计量 MSRVM 和 TMSRVM 随着噪声水平的增加, 先是减小, 然后逐渐增大 (这种特征在表 1-6 也会体现出来). 因此, 在不同的噪声水平下, 表 1-3 中所有的交易信息的模拟数据都是不同的, 也就是在不同噪声水平下真实的波动率矩阵是不一样的.

从表 1-3 我们发现, 随着噪声水平的提高, 估计量 ARVM 和 TARVM 以及估计量 MSRVM 和 TMSRVM 随着噪声水平的增加, 它们的表现是先逐渐减少, 再逐渐增大. 主要是因为低噪声水平下, 噪声对积分波动率矩阵估计影响很小, 尤其当 $\kappa = 0.001$ (相当于无噪声) 时, 而烦琐的降噪运算反而会增加估计量的均方差和标准差. 同时我们发现估计量 MSRVM 和 TMSRVM 在低噪声水平表现不如估计量 ARVM 和 TARVM, 但当噪声水平提高时, 它们的表现优于估计量 ARVM 和 TARVM, 但它们皆呈现逐渐增大的趋势. 而我们的估计量 AERVM 和 TAERVM 则呈现出较强的稳健性, 在不同的噪声水平下表现比其他 4 种估计量都要好.

表 1-3 只资产在不同的 g_1 噪声水平下积分波动率矩阵估计量 MSE 与 se 的比较

噪声水平	ARVM	TARVM	MSRVM	TMSRVM	AERVM	TAERVM
$\kappa = 0.001$	2.163 (0.001)	2.163 (0.001)	3.773 (0.014)	3.773 (0.014)	0.935 (0.031)	0.935 (0.031)
$\kappa = 0.100$	1.482 (0.119)	1.482 (0.119)	3.516 (0.207)	3.516 (0.207)	1.111 (0.040)	1.111 (0.040)
$\kappa = 0.200$	1.071 (0.185)	1.071 (0.185)	3.673 (0.218)	3.673 (0.218)	0.936 (0.041)	0.936 (0.041)
$\kappa = 0.400$	2.349 (0.456)	2.344 (0.452)	4.932 (0.486)	4.930 (0.483)	1.717 (0.035)	1.708 (0.031)
$\kappa = 0.600$	3.687 (0.790)	3.648 (0.771)	6.204 (0.919)	6.199 (0.910)	1.044 (0.026)	1.044 (0.026)
$\kappa = 0.800$	5.723 (1.076)	5.629 (1.006)	7.183 (1.326)	7.154 (1.303)	1.338 (0.043)	1.338 (0.043)
$\kappa = 1.000$	6.334 (1.039)	6.129 (0.916)	7.926 (1.813)	7.822 (1.787)	0.862 (0.025)	0.862 (0.025)
$\kappa = 1.200$	9.065 (1.631)	8.724 (1.619)	8.993 (2.137)	8.886 (2.059)	0.933 (0.033)	0.933 (0.033)
$\kappa = 1.400$	10.444 (2.468)	9.739 (2.737)	10.011 (2.426)	9.543 (2.374)	1.804 (0.041)	1.804 (0.041)
$\kappa = 1.600$	16.158 (2.855)	15.563 (2.662)	13.583 (2.632)	13.083 (2.584)	1.565 (0.022)	1.565 (0.022)

表 2 3 只资产在不同的 g_2 噪声水平下积分波动率矩阵估计量 MSE 与 se 的比较

噪声水平	ARVM	TARVM	MSRVM	TMSRVM	AERVM	TAERVM
$\kappa = 0.001$	2.163 (0.001)	2.163 (0.001)	3.768 (0.014)	3.768 (0.014)	0.928 (0.057)	0.928 (0.057)
$\kappa = 0.100$	1.479 (0.126)	1.479 (0.126)	3.495 (0.202)	3.495 (0.202)	1.104 (0.043)	1.104 (0.043)
$\kappa = 0.200$	1.087 (0.177)	1.087 (0.177)	3.669 (0.211)	3.669 (0.211)	0.937 (0.047)	0.937 (0.047)
$\kappa = 0.400$	2.350 (0.315)	2.339 (0.306)	4.927 (0.479)	4.921 (0.472)	1.721 (0.040)	1.704 (0.031)
$\kappa = 0.600$	3.358 (0.776)	3.546 (0.761)	6.196 (0.902)	6.195 (0.901)	1.050 (0.028)	1.050 (0.028)
$\kappa = 0.800$	5.540 (1.028)	5.429 (0.955)	7.179 (1.304)	7.152 (1.287)	1.338 (0.051)	1.338 (0.051)
$\kappa = 1.000$	6.143 (1.068)	5.937 (0.951)	7.894 (1.810)	7.881 (1.798)	0.863 (0.030)	0.863 (0.030)
$\kappa = 1.200$	8.942 (1.653)	8.784 (1.591)	8.981 (2.102)	8.880 (2.056)	0.932 (0.040)	0.932 (0.040)
$\kappa = 1.400$	9.947 (2.209)	9.280 (2.446)	10.001 (2.395)	9.637 (2.309)	1.807 (0.054)	1.807 (0.054)
$\kappa = 1.600$	15.719 (2.872)	15.154 (2.638)	13.383 (2.637)	12.797 (2.554)	1.562 (0.030)	1.562 (0.030)

表 3 3 只资产在不同的 g_1 噪声水平下积分波动率矩阵估计量 MSE 与 se 的比较

噪声水平	ARVM	TARVM	MSRVM	TMSRVM	AERVM	TAERVM
$\kappa = 0.001$	2.163 (0.001)	2.163 (0.001)	3.773 (0.014)	3.773 (0.014)	0.935 (0.031)	0.935 (0.031)
$\kappa = 0.100$	1.479 (0.126)	1.479 (0.126)	3.512 (0.206)	3.512 (0.206)	1.096 (0.027)	1.096 (0.027)
$\kappa = 0.200$	1.072 (0.185)	1.072 (0.185)	3.672 (0.217)	3.672 (0.217)	0.947 (0.053)	0.947 (0.053)
$\kappa = 0.400$	2.348 (0.456)	2.344 (0.452)	4.930 (0.481)	4.930 (0.481)	1.718 (0.044)	1.702 (0.035)
$\kappa = 0.600$	3.368 (0.790)	3.364 (0.771)	6.201 (0.917)	6.199 (0.916)	1.050 (0.030)	1.050 (0.030)
$\kappa = 0.800$	5.715 (1.076)	5.622 (1.006)	7.183 (1.310)	7.126 (1.299)	1.344 (0.045)	1.344 (0.045)
$\kappa = 1.000$	6.324 (1.038)	6.120 (0.915)	7.918 (1.805)	7.901 (1.804)	0.856 (0.028)	0.856 (0.028)
$\kappa = 1.200$	9.059 (1.631)	8.726 (1.617)	8.992 (2.136)	8.843 (2.091)	0.931 (0.037)	0.931 (0.037)
$\kappa = 1.400$	10.430 (2.466)	9.725 (2.734)	10.008 (2.413)	9.643 (2.363)	1.789 (0.050)	1.789 (0.050)
$\kappa = 1.600$	16.140 (2.854)	15.546 (2.661)	13.579 (2.621)	12.995 (2.518)	1.564 (0.031)	1.564 (0.031)

表 4 50 只资产在不同的 g_1 噪声水平下积分波动率矩阵估计量 MSE 与 se 的比较

噪声水平	ARVM	TARVM	MSRVM	TMSRVM	AERVM	TAERVM
$\kappa = 0.001$	10.955 (0.001)	5.901 (0.001)	13.836 (0.014)	6.288 (0.014)	11.913 (0.070)	4.604 (0.038)
$\kappa = 0.100$	11.990 (0.069)	5.716 (0.088)	15.274 (0.126)	6.074 (0.145)	13.193 (0.017)	5.336 (0.024)
$\kappa = 0.200$	10.450 (0.349)	4.536 (0.064)	13.673 (0.518)	5.289 (0.129)	10.131 (0.021)	4.432 (0.002)
$\kappa = 0.400$	12.613 (0.700)	5.681 (0.043)	14.211 (0.726)	5.667 (0.125)	9.972 (0.034)	4.534 (0.004)
$\kappa = 0.600$	18.425 (0.443)	7.487 (0.104)	18.605 (0.805)	6.932 (0.279)	10.082 (0.012)	4.820 (0.005)
$\kappa = 0.800$	22.080 (0.119)	9.059 (0.326)	22.427 (0.333)	8.945 (0.759)	12.144 (0.074)	4.661 (0.021)
$\kappa = 1.000$	24.867 (0.752)	10.339 (0.425)	24.011 (1.648)	9.567 (0.077)	11.649 (0.021)	4.829 (0.005)
$\kappa = 1.200$	30.313 (1.185)	14.792 (1.792)	29.189 (0.687)	13.466 (0.288)	10.230 (0.026)	5.042 (0.007)
$\kappa = 1.400$	38.770 (1.993)	15.833 (0.397)	36.938 (1.125)	14.719 (0.587)	11.196 (0.041)	4.438 (0.014)
$\kappa = 1.600$	46.655 (2.995)	21.291 (0.406)	46.360 (1.447)	20.540 (0.546)	11.533 (0.020)	4.960 (0.004)

此外, 由于资产数只有 3 只, 所以, 表 1–3 中积分波动率矩阵估计量的稀疏性作用不明显, 截断作用几乎没发挥作用. 下面将展开对更多资产的市场波动率矩阵的估计, 由于资产数的增加, 对波动率

表 5 50 只资产在不同的 g_2 噪声水平下积分波动率矩阵估计量 MSE 与 se 的比较

噪声水平	ARVM	TARVM	MSRVM	TMSRVM	AERVM	TAERVM
$\kappa = 0.001$	10.955 (0.001)	5.901 (0.001)	13.836 (0.014)	6.288 (0.014)	11.850 (0.128)	4.606 (0.024)
$\kappa = 0.100$	11.977 (0.074)	5.717 (0.082)	15.258 (0.101)	6.062 (0.197)	13.147 (0.114)	5.319 (0.031)
$\kappa = 0.200$	10.406 (0.349)	4.579 (0.103)	13.476 (0.493)	5.178 (0.106)	10.098 (0.011)	4.418 (0.008)
$\kappa = 0.400$	12.528 (0.697)	5.632 (0.078)	14.059 (0.726)	5.592 (0.194)	9.950 (0.013)	4.536 (0.009)
$\kappa = 0.600$	18.268 (0.408)	7.357 (0.044)	18.412 (0.733)	6.923 (0.221)	10.059 (0.012)	4.829 (0.032)
$\kappa = 0.800$	21.761 (0.082)	8.963 (0.320)	22.266 (0.496)	8.892 (0.655)	12.152 (0.046)	4.662 (0.014)
$\kappa = 1.000$	24.489 (0.889)	10.077 (0.400)	23.626 (1.511)	9.263 (0.067)	11.626 (0.002)	4.822 (0.004)
$\kappa = 1.200$	29.995 (1.448)	14.867 (1.325)	28.972 (0.687)	13.627 (0.098)	10.232 (0.024)	5.033 (0.038)
$\kappa = 1.400$	37.719 (2.162)	15.299 (0.206)	36.913 (1.117)	14.048 (0.250)	11.216 (0.041)	4.426 (0.025)
$\kappa = 1.600$	45.470 (2.367)	21.407 (0.064)	45.393 (1.531)	20.127 (0.203)	11.526 (0.023)	4.955 (0.003)

表 6 50 只资产在不同的 g_3 噪声水平下积分波动率矩阵估计量 MSE 与 se 的比较

噪声水平	ARVM	TARVM	MSRVM	TMSRVM	AERVM	TAERVM
$\kappa = 0.001$	10.955 (0.001)	5.901 (0.001)	13.836 (0.014)	6.288 (0.014)	11.876 (0.059)	4.606 (0.036)
$\kappa = 0.100$	11.990 (0.069)	5.716 (0.088)	15.274 (0.126)	6.074 (0.145)	13.205 (0.040)	5.319 (0.045)
$\kappa = 0.200$	10.449 (0.349)	4.536 (0.063)	13.669 (0.502)	5.287 (0.127)	10.107 (0.051)	4.427 (0.009)
$\kappa = 0.400$	12.611 (0.699)	5.679 (0.043)	14.208 (0.718)	5.657 (0.123)	9.955 (0.034)	4.534 (0.005)
$\kappa = 0.600$	18.420 (0.443)	7.482 (0.104)	18.600 (0.805)	6.932 (0.273)	10.093 (0.011)	4.823 (0.016)
$\kappa = 0.800$	22.070 (0.120)	9.051 (0.326)	22.418 (0.333)	8.920 (0.758)	12.099 (0.071)	4.669 (0.018)
$\kappa = 1.000$	24.853 (0.752)	10.329 (0.425)	23.997 (1.647)	9.556 (0.076)	11.621 (0.019)	4.831 (0.035)
$\kappa = 1.200$	30.295 (1.185)	14.792 (1.792)	29.172 (0.686)	13.474 (0.287)	10.222 (0.025)	5.046 (0.017)
$\kappa = 1.400$	38.747 (1.993)	15.818 (0.397)	36.914 (1.125)	14.704 (0.584)	11.153 (0.051)	4.400 (0.012)
$\kappa = 1.600$	46.625 (2.993)	21.273 (0.407)	46.328 (1.447)	20.520 (0.547)	11.548 (0.035)	4.960 (0.004)

矩阵估计的难度增加, 所以同时需要考虑波动率矩阵的稀疏性.

表 4-6 给出了 $p = 50$ 个资产在三种不同的噪声模型和不同噪声水平下 6 种估计量的均方差和标准差. 同表 1-3 相比, 随着资产数目的增加, 相应 6 种估计量的均方差和标准差都有所增加, 但整体反映的特征与表 1-3 一致, 都说明考虑带有市场信息的微观结构噪声能够提高估计的效率.

从表 1-6 可以看出, 在本文所提到的 6 种估计量当中, 尤其当资产数很大时, 截断后的估计量通常都比没截断的估计量要更好. 这主要是因为截断估计考虑了稀疏性, 使得该估计量为高维积分波动率矩阵的相合估计, 有效地提高了估计精度.

无论资产数是 3 只还是 50 只, 估计量 AERVM 和 TAERVM 在三种不同的噪声模型和不同噪声水平下表现都是非常突出的. 由此可见, 估计量 ARVM 和 TARVM 以及估计量 MSRVM 和 TMSRVM 不适合估计含有交易信息的噪声 ($\kappa > 0.6$) 的积分波动率矩阵, 这主要是因为它们没有考虑除资产价格外其他的交易信息, 仅简单地认为资产价格只受白噪声影响. 然而市场中的数据都是息息相关的, 往往更多的数据会提供更多的信息帮助我们去估计, 因此, 本文所获得的估计比其他已存在的波动率矩阵更加精确.

关于高维积分波动率矩阵估计, 还有一个很重要的衡量标准就是高维矩阵的特征值. 特征值能反映出起主导作用的因子, 能够刻画估计量的好坏. 在此, 我们仅给出在带有市场信息的微观结构噪声 g_1

下 6 种估计量的特征值情形, 在其他两种带有市场信息的微观结构噪声情形下估计量的表现是一致的, 在此不再赘述, 具体见图 1-5. 由于 50 个资产的特征值有 50 个, 比较 6 种估计量的所有特征值, 图片看起来很不清楚, 再加上起主导作用的资产通常都是排列前几位的, 因此, 图 1-5 给出的是 50 个资产高频数据在 10 种不同的噪声水平下对积分波动率矩阵估计的前 10 个特征值的比较.

从图 1-5 整体来看不难发现估计量 ARVM、MSRVM 和 AERVM 在未做截断处理前, 它们的起主导作用的最大特征值普遍比真实的波动率矩阵的最大特征值大, 而且随着噪声水平的提高, 偏大的程度也越来越严重. 然而经过截断处理后得到的 TARVM、TMSRVM 和 TAERVM, 它们的这种差距减小很多. 同时我们发现, 在低噪声水平 $\kappa \leq 0.4$ 下, 估计量 TARVM 比估计量 TMSRVM 要好, 尤其当 $\kappa = 0.4$ 时, 估计量 TARVM 的特征值曲线达到最优, 说明估计量 TARVM 比估计量 TMSRVM 能更好地解释起主导因素的因子. 然而当噪声水平 $\kappa > 0.4$ 时, 估计量 TARVM 就比不上估计量 TMSRVM, 这个结论与 Tao 等^[16] 的结论基本一致.

随着噪声水平的提高, 估计量 TAERVM 自始至终都表现的很稳定, 没有发生大的起伏, 然而估计量 TARVM 和 TMSRVM 的特征值却随着噪声水平的增大而与真实的波动率矩阵的特征值的差距在不断地增大, 误差也越来越明显, 远劣于本文所提出的估计量 TAERVM 的特征值对主导因子的刻画. 因为本文所提出的估计量的特征值与真实值有更小的偏差, 它与真实波动率矩阵特征值更加接近, 并且均方差也比现有的方法更小. 尤其当 $\kappa \geq 1.0$ 时, 估计量 TARVM 和 TMSRVM 几乎就不合适用来作为波动率矩阵的估计量了, 这主要是由于这些方法没有考虑带有市场信息的微观结构噪声所导致更大偏差. 估计量 MSRVM、TMSRVM、ARVM 和 TARVM 只适用于均值为 0、方差不同的白噪声, 当

图 1 (网络版彩图) 50 只资产在噪声为 g_1 下 6 种估计量前 10 个特征值与真实特征值的比较, (a) 噪声水平 = 0.001; (b) 噪声水平 = 0.1. 黑色实线是真实波动率矩阵的特征值曲线; 红色短虚线是估计量 ARVM 的特征值曲线; 绿色点线是通过门限技巧截断的估计量 TARVM 的特征值曲线; 蓝色点短虚线表示的是 MSRVM 估计量的特征值曲线; 天蓝色长虚线表示的是 TMSRVM 估计量对应的特征值曲线; 剩下的两条是本文提出的估计量 AERVM (玫红色点长虚线) 和估计量 TAERVM (黄色短虚线) 的特征值曲线

图 2 (网络版彩图) 50 只资产在噪声为估计量下 6 种估计量前 10 个特征值与真实特征值的比较, (a) 噪声水平 = 0.2; (b) 噪声水平 = 0.4

图 3 (网络版彩图) 50 只资产在噪声为 g_1 下 6 种估计量前 10 个特征值与真实特征值的比较, (a) 噪声水平 = 0.6; (b) 噪声水平 = 0.8

图 4 (网络版彩图) 50 只资产在噪声为 g_1 下 6 种估计量前 10 个特征值与真实特征值的比较, (a) 噪声水平 = 1.0; (b) 噪声水平 = 1.2

图 5 (网络版彩图) 50 只资产在噪声为 g_1 下 6 种估计量前 10 个特征值与真实特征值的比较, (a) 噪声水平 = 1.4; (b) 噪声水平 = 1.6

微观噪声受新的重大市场信息影响时, 这些估计量无法做出相应的调整, 因此, 这些估计量的特征值曲线与真实的波动率矩阵的特征值曲线随噪声水平的上升表现差异巨大.

6 实证研究

为了考察高维积分波动率矩阵估计量的行为, 本节将利用 20 只上海证券指数²⁾ (包含上证指数) 在 2013 年 1 月 4 日和 1 月 8 日两天的以 5 秒记录的交易数据研究各种积分波动率矩阵估计量的特征值, 由于没有真实的波动率矩阵作对比, 实证研究中我们无法与 Tao 等^[16] 提出的 MSRV 和 TMSRV 估计量进行比较, 所以在此仅考虑估计量 ARVM、TARVM 和本文估计量特征值的比较. 由于上证指数数据记录的非同步性, 所以两天的 20 只资产的总体样本量也是有差异的, 分别为 55,660 和 56,100; 同时为保证两天起主导作用的因子处于一个量级, 我们取 K 值分别为 10 和 15. 为了保证稀疏后起主导作用的因子至少能达到总体的 25%, 两天的门限值 ϖ 定为 5, 具体见表 7 和 8.

从表 7 和 8 可以看出本文提出的估计量起主导作用的特征值小于已存在的估计量起主导作用的特征值, 这一特征与模拟结果一致; 同时本文提出的估计量 AERVM 的特征值没有负数, 保证了积分

表 7 2013 年 1 月 4 日 20 只上证指数不同估计量的特征值比较

			g_1	g_1	g_2	g_2	g_3	g_3
	ARVM	TARVM	$\hat{\Gamma}$	$\mathcal{T}_{\varpi}[\hat{\Gamma}]$	$\hat{\Gamma}$	$\mathcal{T}_{\varpi}[\hat{\Gamma}]$	$\hat{\Gamma}$	$\mathcal{T}_{\varpi}[\hat{\Gamma}]$
1	82.81	62.34	79.31	57.04	79.33	57.19	79.29	57.15
2	8.54	14.75	7.93	13.87	7.94	14.19	7.93	14.18
3	5.48	11.08	5.61	10.93	5.55	10.50	5.59	10.50
4	3.08	10.47	3.29	8.38	3.29	9.22	3.29	9.22
5	2.56	7.19	2.69	6.59	2.69	6.65	2.69	6.65
6	2.31	6.77	2.41	6.40	2.40	6.57	2.40	6.56
7	2.17	5.17	2.29	5.97	2.29	5.95	2.29	5.95
8	1.64	4.20	1.82	4.16	1.82	4.15	1.82	4.16
9	1.58	3.34	1.75	3.92	1.75	3.75	1.75	3.75
10	1.31	1.54	1.41	2.15	1.41	2.13	1.42	2.13
11	0.81	0.16	1.01	0.53	1.02	0.28	1.02	0.28
12	0.54	0.00	0.85	0.00	0.85	0.00	0.85	0.00
13	0.47	-0.00	0.62	-0.00	0.62	0.00	0.62	-0.00
14	0.21	-0.08	0.49	-0.13	0.49	-0.70	0.49	-0.69
15	0.13	-0.89	0.45	-0.86	0.45	-0.86	0.45	-0.85
16	0.06	-2.20	0.26	-1.29	0.26	-1.24	0.26	-1.24
17	-0.03	-4.04	0.23	-1.93	0.23	-1.99	0.23	-1.99
18	-0.10	-5.37	0.09	-4.07	0.09	-3.98	0.09	-3.97
19	-0.21	-5.96	0.05	-5.40	0.05	-5.41	0.05	-5.41
20	-0.21	-9.28	0.04	-7.45	0.04	-7.62	0.04	-7.62

2) 20 只上海证券指数: 上证材料、上证地企、上证工业、上证公用、上证海外、上证沪企、上证金融、上证可选、上证全指、上证龙头、上证民企、上证能源、上证消费、上证新兴、上证商品、上证信息、上证央企、上证国企、上证周期、上证指数.

表 8 2013 年 1 月 8 日 20 只上证指数不同估计量的特征值比较

	ARVM	TARVM	g_1 $\hat{\Gamma}$	g_1 $\tau_{\infty}[\hat{\Gamma}]$	g_2 $\hat{\Gamma}$	g_2 $\tau_{\infty}[\hat{\Gamma}]$	g_3 $\hat{\Gamma}$	g_3 $\tau_{\infty}[\hat{\Gamma}]$
1	101.98	74.25	95.67	61.41	95.62	61.39	95.67	61.41
2	10.53	19.20	9.97	16.57	9.97	16.57	9.97	16.57
3	4.57	14.59	4.23	13.00	4.22	12.99	4.23	13.00
4	3.70	10.85	3.64	9.52	3.65	9.51	3.64	9.52
5	3.44	9.09	3.23	7.93	3.23	7.92	3.23	7.93
6	2.57	8.07	2.59	7.77	2.59	7.75	2.59	7.77
7	2.13	7.87	2.16	6.95	2.16	6.95	2.16	6.96
8	1.96	7.24	1.97	6.68	1.97	6.67	1.97	6.68
9	1.68	5.95	1.77	6.54	1.76	6.53	1.77	6.54
10	1.55	4.30	1.58	6.01	1.58	6.01	1.58	6.02
11	1.22	2.34	1.42	5.18	1.42	5.16	1.42	5.18
12	0.92	0.76	0.97	3.73	0.97	3.73	0.97	3.73
13	0.83	0.22	0.96	3.32	0.96	3.32	0.96	3.32
14	0.38	-0.05	0.57	0.60	0.57	0.60	0.57	0.60
15	0.25	-1.13	0.46	-0.09	0.46	-0.09	0.46	-0.09
16	0.10	-1.43	0.34	-0.76	0.34	-0.75	0.34	-0.76
17	0.08	-2.31	0.21	-2.73	0.21	-2.72	0.21	-2.73
18	0.00	-3.93	0.17	-5.50	0.17	-5.49	0.17	-5.50
19	-0.05	-7.53	0.06	-6.06	0.06	-6.06	0.06	-6.06
20	-0.11	-14.99	0.04	-12.11	0.04	-12.10	0.04	-12.10

波动率矩阵的半正定性, 符合积分波动率矩阵的特点. 然而, 估计量 ARVM、TARVM 和 TAERVM 中存在负的特征值, 这将无法保证波动率矩阵的半正定性, 需要一些技术的处理, 无疑会增加误差; 同时也说明本文的截断方法无法保证原有估计量的半正定性. 此外, 表 7 和 8 中利用实际数据获得估计的最大特征值与次最大特征值悬殊很大, 而模拟中不存在这个问题. 这主要是因为模拟中, 资产的对数价格的模型设置是随机的, 没有个性特征, 从而 p 个资产的地位是等价的, 积分波动率矩阵的特征值大小差异不明显. 然而现实中的资产不仅具有自身的特性, 而且与其他资产可能存在某种正向或反向的内在联系, 通过计算积分波动率矩阵就会凸显出那些在市场上占主导地位资产的作用. 众所周知, 上证指数包含上海各个行业的指数信息, 起绝对主导性作用, 因此, 表 7 和 8 中所有估计量的最大特征值与其他特征值会出现如此悬殊的特征.

尽管总体来看本文的估计量特征值在实际应用中表现的没有模拟研究中那么突出, 但这并不完全是估计量本身的原因, 而是因为市场微观结构噪声建模过于简单, 无法真实刻画现实市场的微观结构噪声. 这无疑对市场微观结构噪声建模提出了新的挑战.

7 结论

本文根据实际情况, 在带有市场信息的微观结构噪声模型下, 提出了高维积分波动率矩阵的估计

AERVM 和 TAERVM; 从理论上证明了在一定的假设下本文所提出的估计量是相合的, 且逐点收敛速度达到样本量 $n^{-1/3}$, 优于目前已有的高维积分波动率矩阵估计量. 从模拟中, 我们进一步证实了本文所提出的估计量优于已存在的高维积分波动率矩阵估计量. 在实证研究中, 本文所提出的估计量稍微优于已存在的估计量, 这可能是由于本文对微观市场结构噪声建模过于简化, 更多重要的市场微观信息没有被发掘和应用, 需要进一步研究, 寻找重要信息进一步建模. 此外, 本文没有进一步研究 BAERVM, 是因为市场资产是无序的, 很难保证波动率矩阵满足退化条件. 所以, 在模拟和实证中, 本文没有对其进一步阐述.

事实上, 本文所提出的高维积分波动率矩阵估计量可以进一步推广, 从 K 出发, 类似于估计量 MSRV 推广估计量 ARVM, 得到 MSERVM 和 TMSERVM. 虽然效果可能会更好, 但收敛速度不会进一步提高, 有致力此方面研究的学者可以深入探讨. 另一方面, 感兴趣的学者可以从截断方法出发, 提出新的截断方法保证原有估计量的半正定性.

8 定理的证明

8.1 定理 4.1 的证明

记 C 是一般的常数, 它的值与 n 和 p 无关, 并且其随着不同的情况改变, 但仍为常数. 当 n 和 p 趋向于无穷时, O_p 和 o_p 分别记为概率意义上的同阶和高阶.

8.1.1 $\hat{\Gamma}$ 的分解

记 $\mathbf{Y}_r^k = (Y_1(\tau_{1,r}^k), \dots, Y_p(\tau_{p,r}^k))^T$, $\mathbf{X}_r^k = (X_1(\tau_{1,r}^k), \dots, X_p(\tau_{p,r}^k))^T$ 和 $\mathbf{g}_r^k = (g(\mathbf{Z}_1(\tau_{1,r}^k); \boldsymbol{\theta}_1), \dots, g(\mathbf{Z}_p(\tau_{p,r}^k); \boldsymbol{\theta}_p))^T$ 分别为在给定时间 τ_r 下 p 个资产的观察的对数价格向量、真实对数价格向量和噪声向量. 注意到由于 $\tau_{1,r}^k, \dots, \tau_{p,r}^k$ 是不等的, 所以从某种意义上说这些向量在不同的时间点上是非同步的. 因此, 需要对 $\hat{\Gamma}$ 进行适当地分解.

$$\begin{aligned}\hat{\Gamma} &= \frac{1}{K} \sum_{k=1}^K \sum_{r=1}^m (\Delta \widehat{\mathbf{X}}_r^k) (\Delta \widehat{\mathbf{X}}_r^k)^T = \frac{1}{K} \sum_{k=1}^K \sum_{r=1}^m (\Delta \mathbf{Y}_r^k - \Delta \widehat{\mathbf{g}}_r^k) (\Delta \mathbf{Y}_r^k - \Delta \widehat{\mathbf{g}}_r^k)^T \\ &= \frac{1}{K} \sum_{k=1}^K \sum_{r=1}^m (\Delta \mathbf{X}_r^k + \Delta \mathbf{g}_r^k - \Delta \widehat{\mathbf{g}}_r^k) (\Delta \mathbf{X}_r^k + \Delta \mathbf{g}_r^k - \Delta \widehat{\mathbf{g}}_r^k)^T \\ &= \frac{1}{K} \sum_{k=1}^K \sum_{r=1}^m (\Delta \mathbf{X}_r^k) (\Delta \mathbf{X}_r^k)^T + \frac{1}{K} \sum_{k=1}^K \sum_{r=1}^m (\Delta \widehat{\mathbf{g}}_r^k - \Delta \mathbf{g}_r^k) (\Delta \widehat{\mathbf{g}}_r^k - \Delta \mathbf{g}_r^k)^T \\ &\quad - \frac{1}{K} \sum_{k=1}^K \sum_{r=1}^m (\Delta \mathbf{X}_r^k) (\Delta \widehat{\mathbf{g}}_r^k - \Delta \mathbf{g}_r^k)^T - \frac{1}{K} \sum_{k=1}^K \sum_{r=1}^m (\Delta \widehat{\mathbf{g}}_r^k - \Delta \mathbf{g}_r^k) (\Delta \mathbf{X}_r^k)^T \\ &=: \mathbf{I}^{(1)} + \mathbf{I}^{(2)} + \mathbf{I}^{(3)} + \mathbf{I}^{(4)},\end{aligned}$$

其中 $\mathbf{I}^{(1)}$ 与噪声项 $\mathbf{g}_r^k$ 无关, 相当于无噪声非同步真实价格对数 $\mathbf{X}_r^k$ 的平均实波动率矩阵估计量. 于是, 首先考虑 $\mathbf{I}^{(1)}$, 为计算方便, 我们需要将其进一步分解. 由假设 4.3 易知 $\tau_{i,r-1}^k \leq \tau_{r-1}^k < \tau_{i,r}^k \leq \tau_r^k$, 我们可以在 $\mathbf{X}_r^k$ 和 $\mathbf{X}_{r-1}^k$ 之间插入同步真实对数价格 $\mathbf{X}(\tau_r^k)$ 和 $\mathbf{X}(\tau_{r-1}^k)$, 形如

$$\mathbf{X}_r^k - \mathbf{X}_{r-1}^k = [\mathbf{X}_r^k - \mathbf{X}(\tau_r^k)] + [\mathbf{X}(\tau_r^k) - \mathbf{X}(\tau_{r-1}^k)] + [\mathbf{X}(\tau_{r-1}^k) - \mathbf{X}_{r-1}^k].$$

利用上述表达式对 $\mathbf{I}^{(1)}$ 进行展开, 可得

$$\begin{aligned}\mathbf{I}^{(1)} &= \frac{1}{K} \sum_{k=1}^K \sum_{r=1}^m (\mathbf{X}_r^k - \mathbf{X}_{r-1}^k)(\mathbf{X}_r^k - \mathbf{X}_{r-1}^k)^T \\ &= \frac{1}{K} \sum_{k=1}^K \sum_{r=1}^m \{ [\mathbf{X}(\tau_r^k) - \mathbf{X}(\tau_{r-1}^k)][\mathbf{X}(\tau_r^k) - \mathbf{X}(\tau_{r-1}^k)]^T \\ &\quad + [\mathbf{X}_r^k - \mathbf{X}(\tau_r^k)][\mathbf{X}_r^k - \mathbf{X}(\tau_r^k)]^T + [\mathbf{X}_r^k - \mathbf{X}(\tau_r^k)][\mathbf{X}(\tau_r^k) - \mathbf{X}(\tau_{r-1}^k)]^T \\ &\quad + [\mathbf{X}_r^k - \mathbf{X}(\tau_r^k)][\mathbf{X}(\tau_{r-1}^k) - \mathbf{X}_{r-1}^k]^T + [\mathbf{X}(\tau_{r-1}^k) - \mathbf{X}_{r-1}^k][\mathbf{X}_r^k - \mathbf{X}(\tau_r^k)]^T \\ &\quad + [\mathbf{X}(\tau_{r-1}^k) - \mathbf{X}_{r-1}^k][\mathbf{X}(\tau_r^k) - \mathbf{X}(\tau_{r-1}^k)]^T + [\mathbf{X}(\tau_{r-1}^k) - \mathbf{X}_{r-1}^k][\mathbf{X}(\tau_{r-1}^k) - \mathbf{X}_{r-1}^k]^T \\ &\quad + [\mathbf{X}(\tau_r^k) - \mathbf{X}(\tau_{r-1}^k)][\mathbf{X}_r^k - \mathbf{X}(\tau_r^k)]^T + [\mathbf{X}(\tau_r^k) - \mathbf{X}(\tau_{r-1}^k)][\mathbf{X}(\tau_{r-1}^k) - \mathbf{X}_{r-1}^k]^T \} \\ &= \mathbf{V} + \mathbf{H}(1) + \cdots + \mathbf{H}(8),\end{aligned}$$

其中 $\mathbf{V}$ 是同步真实对数价格 $\mathbf{X}(\tau_r^k)$ 的平均实波动率估计量, $\mathbf{H}(1), \dots, \mathbf{H}(8)$ 来自于真实对数价格的非同步性. 综合上面的分解, 我们得到

$$\hat{\Gamma} - \Gamma = (\mathbf{V} - \Gamma) + \mathbf{H}(1) + \cdots + \mathbf{H}(8) + \mathbf{I}^{(2)} + \mathbf{I}^{(3)} + \mathbf{I}^{(4)}. \quad (8.1)$$

下一小节将逐个考虑它们的阶数.

8.1.2 逐项分析

因为我们的目的是证明 $E(|\hat{\Gamma}_{ij} - \Gamma_{ij}|^\beta) \leq Cn^{-\frac{\beta}{2}}$, 考虑的是矩阵每个元素的关系, 因此, 我们分析 (8.1) 中的每项矩阵的每个元素. 已知 $\mathbf{I}^{(1)} = (I_{ij}^{(1)})_{0 \leq i, j \leq p}$, 则 $\mathbf{I}^{(1)} - \Gamma = (I_{ij}^{(1)} - \Gamma_{ij})_{0 \leq i, j \leq p}$. 对于 $\mathbf{V} - \Gamma, \mathbf{H}(1), \dots, \mathbf{H}(8)$ 9 项矩阵中每个元素的阶数参见文献 [14, 定理 1] 的证明过程, 容易得到 $E|I_{ij}^{(1)} - \Gamma_{ij}|^\beta \leq C[m^{-\frac{\beta}{2}} + (\frac{m}{n})^\beta + n^{-\frac{\beta}{2}}]$.

下面考虑与噪声项 g_r^k 有关的部分, 即矩阵 $\mathbf{I}^{(2)}, \mathbf{I}^{(3)}$ 和 $\mathbf{I}^{(4)}$ 中的元素. 首先来看 $\mathbf{I}^{(2)}$ 中第 (i, j) 个元素:

$$I_{ij}^{(2)} = \frac{1}{K} \sum_{k=1}^K \sum_{r=1}^m [\Delta g(\mathbf{Z}_i(\tau_{i,r}^k); \hat{\boldsymbol{\theta}}_i) - \Delta g(\mathbf{Z}_i(\tau_{i,r}^k); \boldsymbol{\theta}_i)][\Delta g(\mathbf{Z}_j(\tau_{j,r}^k); \hat{\boldsymbol{\theta}}_j) - \Delta g(\mathbf{Z}_j(\tau_{j,r}^k); \boldsymbol{\theta}_j)]. \quad (8.2)$$

根据假设 4.1, 由文献 [25, 定理 1] 知, $\hat{\boldsymbol{\theta}}_i$ 是 $\boldsymbol{\theta}_i$ 的相合估计, 且 $\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i = O_p(\frac{1}{n})$. 于是, 对 $\forall \beta \geq 2$, 由 (8.2) 和假设 4.1 得

$$\begin{aligned}|I_{ij}^{(2)}|^\beta &\leq \left\{ \frac{1}{K} \sum_{k=1}^K \sum_{r=1}^m [L_0(\mathbf{Z}_i(\tau_{i,r}^k)) + L_0(\mathbf{Z}_i(\tau_{i,r-1}^k))] |\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i| [L_0(\mathbf{Z}_j(\tau_{j,r}^k)) + L_0(\mathbf{Z}_j(\tau_{j,r-1}^k))] |\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j| \right\}^\beta \\ &\leq \frac{1}{K} \sum_{k=1}^K m^{\beta-1} \sum_{r=1}^m \{ [L_0(\mathbf{Z}_i(\tau_{i,r}^k)) + L_0(\mathbf{Z}_i(\tau_{i,r-1}^k))]^\beta |\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i|^\beta \\ &\quad \times [L_0(\mathbf{Z}_j(\tau_{j,r}^k)) + L_0(\mathbf{Z}_j(\tau_{j,r-1}^k))]^\beta |\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j|^\beta \} \\ &\leq m^\beta O_p\left(\frac{1}{n^{2\beta}}\right) \max_{1 \leq r \leq m, 1 \leq k \leq K} \{ [L_0(\mathbf{Z}_i(\tau_{i,r}^k)) + L_0(\mathbf{Z}_i(\tau_{i,r-1}^k))]^\beta [L_0(\mathbf{Z}_j(\tau_{j,r}^k)) + L_0(\mathbf{Z}_j(\tau_{j,r-1}^k))]^\beta \}.\end{aligned}$$

上式的第一个不等号由假设 4.1(v) 得到; 第二个不等号由凸函数性质得到. 因此可以获得

$$E|I_{ij}^{(2)}|^\beta \leq C m^\beta n^{-2\beta},$$

对于矩阵 $\mathbf{I}^{(3)}$ 和 $\mathbf{I}^{(4)}$, 它们是同阶的, 所以只考虑矩阵 $\mathbf{I}^{(3)}$ 中的元素,

$$\begin{aligned} |I_{ij}^{(3)}|^\beta &= \left| \frac{1}{K} \sum_{k=1}^K \sum_{r=1}^m [\Delta X_i(\tau_{i,r}^k)] [\Delta g(\mathbf{Z}_j(\tau_{j,r}^k); \hat{\boldsymbol{\theta}}_j) - \Delta g(\mathbf{Z}_j(\tau_{j,r}^k); \boldsymbol{\theta}_j)] \right|^\beta \\ &\leq \frac{1}{K} \sum_{k=1}^K m^{\beta-1} \sum_{r=1}^m [(\Delta X_i(\tau_{i,r}^k))^2]^{\frac{\beta}{2}} [L_0(\mathbf{Z}_j(\tau_{j,r}^k)) + L_0(\mathbf{Z}_j(\tau_{j,r-1}^k))]^\beta |\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j|^\beta \\ &\leq C \left\{ \frac{1}{K} \sum_{k=1}^K \sum_{r=1}^m m^{\beta-1} [(\Delta X_i(\tau_{i,r}^k))^2]^{\frac{\beta}{2}} \right\} \max_{1 \leq r \leq m, 1 \leq k \leq K} [L_0(\mathbf{Z}_j(\tau_{j,r}^k)) + L_0(\mathbf{Z}_j(\tau_{j,r-1}^k))]^\beta O_p(n^{-\beta}). \end{aligned}$$

由此不难获得 $E|I_{ij}^{(3)}|^\beta \leq C m^{\beta-1} n^{-\beta} \frac{1}{K} \sum_{k=1}^K \sum_{r=1}^m (\Delta \tau_{i,r}^k)^{\frac{\beta}{2}} \leq C m^\beta n^{-\frac{3}{2}\beta}$.

综上所述, 可以证明

$$\begin{aligned} E|\hat{\Gamma}_{ij} - \Gamma_{ij}|^\beta &\leq C \{E|I_{ij}^{(1)} - \Gamma_{ij}|^\beta + E|I_{ij}^{(2)}|^\beta + E|I_{ij}^{(3)}|^\beta + E|I_{ij}^{(4)}|^\beta\} \\ &\leq C \left[m^{-\frac{\beta}{2}} + \left(\frac{m}{n} \right)^\beta + n^{-\frac{\beta}{2}} + m^\beta n^{-2\beta} + 2m^\beta n^{-\frac{3}{2}\beta} \right]. \end{aligned}$$

取 $m = O(n^{\frac{2}{3}})$, 则 $K = O(n^{\frac{1}{3}})$, 从而得到 $E|\hat{\Gamma}_{ij} - \Gamma_{ij}|^\beta \leq C n^{-\frac{\beta}{3}}$, 其中 $\beta \geq 2$. 当 $\beta = 1$ 时, $E|\hat{\Gamma}_{ij} - \Gamma_{ij}| \leq (E|\hat{\Gamma}_{ij} - \Gamma_{ij}|^\beta)^{\frac{1}{\beta}} \leq C n^{-\frac{1}{3}}$, 从而定理 4.1 得证.

8.2 定理 4.2 的证明

由 l_d -范数定义及其三角不等式得

$$\|\mathcal{B}_b[\hat{\Gamma}] - \Gamma\|_2 \leq \|\mathcal{B}_b[\hat{\Gamma}] - \Gamma\|_\infty \leq \|\mathcal{B}_b[\hat{\Gamma}] - \mathcal{B}_b[\Gamma]\|_\infty + \|\mathcal{B}_b[\Gamma] - \Gamma\|_\infty. \quad (8.3)$$

注意到

$$\begin{aligned} \|\mathcal{B}_b[\Gamma] - \Gamma\|_\infty &\leq \max_{1 \leq i \leq p} \sum_{|i-j| > b} |\Gamma_{ij}| \leq 2 \max_{1 \leq i \leq p} \sum_{k=[b]+1}^{\infty} |\Gamma_{i,k-i}| \\ &\leq 2 \sum_{k=[b]+1}^{\infty} \frac{M}{1+k^{\alpha+1}} < 2M \sum_{k=[b]+1}^{\infty} k^{-(\alpha+1)} \leq \frac{2M}{\alpha} b^{-\alpha}, \end{aligned} \quad (8.4)$$

其中 $[b]$ 表示向下取整. 接下来考虑 (8.3) 的第一项:

$$\|\mathcal{B}_b[\hat{\Gamma}] - \mathcal{B}_b[\Gamma]\|_\infty \leq \max_{1 \leq i \leq p} \sum_{j=1}^p |\hat{\Gamma}_{ij} - \Gamma_{ij}| I(|i-j| \leq b) \leq (2b+1) \max_{|i-j| \leq b} |\hat{\Gamma}_{ij} - \Gamma_{ij}|. \quad (8.5)$$

对于固定的 i , 有 $\sum_{j=1}^p I(|i-j| \leq b) \leq 2[b] + 1$. 由定理 4.1 的结论 (4.1) 得

$$P\left(\max_{|i-j| \leq b} |\hat{\Gamma}_{ij} - \Gamma_{ij}| > d\right) \leq \sum_{|i-j| \leq b} P(|\hat{\Gamma}_{ij} - \Gamma_{ij}| > d) \leq p \frac{2[b] + 1}{d^\beta} \max_{|i-j| \leq b} E|\hat{\Gamma}_{ij} - \Gamma_{ij}|^\beta \leq \frac{Cpb}{d^\beta n^{\beta/3}}.$$

结合 (8.5), 可得

$$P(\|\mathcal{B}_b[\hat{\Gamma}] - \mathcal{B}_b[\Gamma]\|_\infty > d) \leq P\left(\max_{|i-j| \leq b} |\hat{\Gamma}_{ij} - \Gamma_{ij}| > \frac{d}{2b+1}\right) \leq \frac{Cpb^{\beta+1}}{d^\beta n^{\beta/3}} \leq \frac{Cb^{-\alpha\beta}}{d^\beta}. \quad (8.6)$$

取适当的带宽 b 和充分小的 d 使得 $pb^{\beta+1}n^{-\beta/3} \sim b^{-\alpha\beta}$, $d = 2d_1b^{-\alpha} \sim (pn^{-\beta/3})^{\frac{\alpha-k}{\alpha\beta+\beta+1}}$, $\alpha > k > 0$, 于是得到当 $n \rightarrow \infty$ 时, $d_1 = (pn^{-\beta/3})^{\frac{-k}{\alpha\beta+\beta+1}} \rightarrow \infty$.

结合 (8.3)、(8.4) 和 (8.6), 可得

$$\begin{aligned} P(\|\mathcal{B}_b[\widehat{\Gamma}] - [\Gamma]\|_\infty > d) &\leq P\left(\|\mathcal{B}_b[\widehat{\Gamma}] - \mathcal{B}_b[\Gamma]\|_\infty > \frac{d}{2}\right) + P\left(\|\mathcal{B}_b[\Gamma] - \Gamma\|_\infty > \frac{d}{2}\right) \\ &\leq \frac{C}{d_1^\beta} + P\left(M > \frac{\alpha d_1}{2}\right) \leq \frac{C}{d_1^\beta} + \frac{2E[M]}{\alpha d_1} \rightarrow 0, \quad d_1 \rightarrow \infty. \end{aligned}$$

从而定理 4.2 得证.

8.3 定理 4.3 的证明

为证明定理 4.3, 首先介绍一个引理:

引理 8.1 假设 Γ 满足稀疏条件 (3.7), $\varpi = n^{-1/3}p^{2/\beta}h_{n,p}$, 且 (4.1) 成立, 则

$$(1) \quad \max_{1 \leq i, j \leq p} |\widehat{\Gamma}_{ij} - \Gamma_{ij}| = O_p(\varpi), \quad (8.7)$$

$$(2) \quad \max_{1 \leq i \leq p} \sum_{j=1}^p I(|\Gamma_{ij}| \geq \varpi) = O_p(\pi(p)\varpi^{-\delta}), \quad (8.8)$$

$$(3) \quad \max_{1 \leq i \leq p} \sum_{j=1}^p I(|\widehat{\Gamma}_{ij}| \geq \varpi, |\Gamma_{ij}| < \varpi) = O_p(\pi(p)\varpi^{-\delta}). \quad (8.9)$$

证明 (1) 利用 Markov 不等式和定理 4.1 中 (4.1) 可得

$$P\left(\max_{1 \leq i, j \leq p} |\widehat{\Gamma}_{ij} - \Gamma_{ij}| > \varpi\right) \leq \sum_{0 \leq i, j \leq p} P(|\widehat{\Gamma}_{ij} - \Gamma_{ij}| > \varpi) \leq \frac{Cp^2}{\varpi^\beta n^{\beta/3}} = \frac{C}{h_{n,p}^\beta} \rightarrow 0, \quad n, p \rightarrow \infty.$$

从而结论 (8.7) 得证.

(2) 根据稀疏条件 (3.7), 可得

$$\max_{1 \leq i \leq p} \sum_{j=1}^p I(|\Gamma_{ij}| \geq \varpi) \leq \max_{1 \leq i \leq p} \sum_{j=1}^p \frac{|\Gamma_{ij}|^\delta}{\varpi^\delta} I(|\Gamma_{ij}| \geq \varpi) \leq \varpi^{-\delta} \max_{1 \leq i \leq p} \sum_{j=1}^p |\Gamma_{ij}|^\delta = O_p(\pi(p)\varpi^{-\delta}). \quad (8.10)$$

从而结论 (8.8) 得证.

(3) 利用定理 4.1 中 (4.1), 可得

$$\begin{aligned} P\left(\max_{1 \leq i \leq p} \sum_{j=1}^p I\left\{|\widehat{\Gamma}_{ij} - \Gamma_{ij}| > \frac{\varpi}{2}\right\} > 0\right) &\leq P\left(\max_{1 \leq i, j \leq p} |\widehat{\Gamma}_{ij} - \Gamma_{ij}| > \frac{\varpi}{2}\right) \\ &\leq \frac{Cp^2}{\varpi^\beta n^{\beta/3}} = \frac{C}{h_{n,p}^\beta} \rightarrow 0, \quad n, p \rightarrow \infty. \end{aligned}$$

因此,

$$\max_{1 \leq i \leq p} \sum_{j=1}^p I\left(|\widehat{\Gamma}_{ij} - \Gamma_{ij}| \geq \frac{\varpi}{2}\right) = o_p(1). \quad (8.11)$$

根据 (8.10) 和 (8.11) 可得

$$\begin{aligned}
 & \max_{1 \leq i \leq p} \sum_{j=1}^p I\{|\hat{\Gamma}_{ij}| \geq \varpi, |\Gamma_{ij}| < \varpi\} \\
 & \leq \max_{1 \leq i \leq p} \sum_{j=1}^p I\left\{|\hat{\Gamma}_{ij}| \geq \varpi, |\Gamma_{ij}| \leq \frac{\varpi}{2}\right\} + \max_{1 \leq i \leq p} \sum_{j=1}^p I\left\{|\hat{\Gamma}_{ij}| \geq \varpi, \frac{\varpi}{2} < |\Gamma_{ij}| < \varpi\right\} \\
 & \leq \max_{1 \leq i \leq p} \sum_{j=1}^p I\left\{|\hat{\Gamma}_{ij} - \Gamma_{ij}| \geq \frac{\varpi}{2}\right\} + \max_{1 \leq i \leq p} \sum_{j=1}^p I\left\{|\Gamma_{ij}| > \frac{\varpi}{2}\right\} \\
 & \leq o_p(1) + \left(\frac{\varpi}{2}\right)^{-\delta} M\pi(p) = O_p(\pi(p)\varpi^{-\delta}).
 \end{aligned}$$

从而结论 (8.9) 得证. $\square$

下面开始证明定理 4.3. 由范数的三角不等式易知,

$$\|\mathcal{T}_\varpi[\hat{\Gamma}] - \Gamma\|_2 \leq \|\mathcal{T}_\varpi[\hat{\Gamma}] - \mathcal{T}_\varpi[\Gamma]\|_2 + \|\mathcal{T}_\varpi[\Gamma] - \Gamma\|_2 \leq \|\mathcal{T}_\varpi[\hat{\Gamma}] - \mathcal{T}_\varpi[\Gamma]\|_\infty + \|\mathcal{T}_\varpi[\Gamma] - \Gamma\|_\infty.$$

显然,

$$\|\mathcal{T}_\varpi[\Gamma] - \Gamma\|_\infty = \max_{1 \leq i \leq p} \sum_{j=1}^p |\Gamma_{ij}| I(|\Gamma_{ij}| \leq \varpi) = O_p(\pi(p)\varpi^{1-\delta}). \quad (8.12)$$

根据简单的代数运算有

$$\begin{aligned}
 \|\mathcal{T}_\varpi[\hat{\Gamma}] - \mathcal{T}_\varpi[\Gamma]\|_\infty &= \max_{1 \leq i \leq p} \sum_{j=1}^p |\hat{\Gamma}_{ij} I(|\hat{\Gamma}_{ij}| \geq \varpi) - \Gamma_{ij} I(|\Gamma_{ij}| \geq \varpi)| \\
 &\leq \max_{1 \leq i \leq p} \sum_{j=1}^p |\hat{\Gamma}_{ij} - \Gamma_{ij}| I(|\hat{\Gamma}_{ij}| \geq \varpi, |\Gamma_{ij}| \geq \varpi) \\
 &\quad + \max_{1 \leq i \leq p} \sum_{j=1}^p |\hat{\Gamma}_{ij}| I(|\hat{\Gamma}_{ij}| \geq \varpi, |\Gamma_{ij}| < \varpi) \\
 &\quad + \max_{1 \leq i \leq p} \sum_{j=1}^p |\Gamma_{ij}| I(|\hat{\Gamma}_{ij}| < \varpi, |\Gamma_{ij}| \geq \varpi) \\
 &=: I_1 + I_2 + I_3.
 \end{aligned}$$

注意到

$$\begin{aligned}
 I_1 &= \max_{1 \leq i \leq p} \sum_{j=1}^p |\hat{\Gamma}_{ij} - \Gamma_{ij}| I(|\hat{\Gamma}_{ij}| \geq \varpi, |\Gamma_{ij}| \geq \varpi) \\
 &\leq \max_{1 \leq i, j \leq p} |\hat{\Gamma}_{ij} - \Gamma_{ij}| \max_{1 \leq i \leq p} \sum_{j=1}^p I(|\Gamma_{ij}| \geq \varpi) = O_p(\pi(p)\varpi^{1-\delta}).
 \end{aligned}$$

上式最后一步由引理 8.1 中的 (8.7) 和 (8.8) 获得.

根据三角不等式和引理 8.1 中的 (8.7)、(8.9) 和 (8.12) 易知,

$$I_2 = \max_{1 \leq i \leq p} \sum_{j=1}^p |\hat{\Gamma}_{ij}| I(|\hat{\Gamma}_{ij}| \geq \varpi, |\Gamma_{ij}| < \varpi) \leq \max_{1 \leq i \leq p} \sum_{j=1}^p (|\hat{\Gamma}_{ij} - \Gamma_{ij}| + |\Gamma_{ij}|) I(|\hat{\Gamma}_{ij}| \geq \varpi, |\Gamma_{ij}| < \varpi)$$

$$\begin{aligned} &\leq \max_{1 \leq i, j \leq p} |\hat{\Gamma}_{ij} - \Gamma_{ij}| \sum_{j=1}^p I(|\hat{\Gamma}_{ij}| \geq \varpi, |\Gamma_{ij}| < \varpi) + \max_{1 \leq i \leq p} \sum_{j=1}^p |\Gamma_{ij}| I(|\Gamma_{ij}| < \varpi) \\ &= O_p(\pi(p)\varpi^{1-\delta}). \end{aligned}$$

同 I_2 类似地可得 $I_3 = \max_{1 \leq i \leq p} \sum_{j=1}^p |\Gamma_{ij}| I(|\hat{\Gamma}_{ij}| < \varpi, |\Gamma_{ij}| \geq \varpi) = O_p(\pi(p)\varpi^{1-\delta})$. 因此, 可证明 $\|\mathcal{T}_\varpi[\hat{\Gamma}] - \mathcal{T}_\varpi[\Gamma]\|_\infty = O_p(\pi(p)\varpi^{1-\delta})$.

致谢 作者非常感谢审稿专家提出的宝贵意见和建议.

参考文献

- Andersen T G, Bollerslev T, Diebold F X, et al. Modeling and forecasting realized volatility. *Econometrica*, 2003, 71: 579–625
- Zhang L, Mykland P A, Ait-Sahalia Y. A tale of two time scales. *J Amer Statist Assoc*, 2005, 100: 1394–1411
- Zhang L. Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach. *Bernoulli*, 2006, 12: 1019–1043
- Fan J, Wang Y. Multi-scale jump and volatility analysis for high-frequency financial data. *J Amer Statist Assoc*, 2007, 102: 1349–1362
- Mancino M E, Sanfelici S. Robustness of Fourier estimator of integrated volatility in the presence of microstructure noise. *Comput Statist Data Anal*, 2008, 52: 2966–2989
- Barndorff-Nielsen O E, Hansen P R, Lunde A, et al. Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica*, 2008, 76: 1481–1536
- Jacod J, Li Y, Mykland P A, et al. Microstructure noise in the continuous case: The pre-averaging approach. *Stochastic Process Appl*, 2009, 119: 2249–2276
- Xiu D. Quasi-maximum likelihood estimation of volatility with high frequency data. *J Econometrics*, 2010, 159: 235–250
- Hayashi T, Yoshida N. On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli*, 2005, 11: 359–379
- Zhang L. Estimating covariation: Epps effect, microstructure noise. *J Econometrics*, 2011, 160: 33–47
- Christensen K, Kinnebrock S, Podolskij M. Pre-averaging estimators of the ex-post covariance matrix in noisy diffusion models with non-synchronous data. *J Econometrics*, 2010, 159: 116–133
- Ait-Sahalia Y, Fan J, Xiu D. High-frequency covariance estimates with noisy and asynchronous financial data. *J Amer Statist Assoc*, 2010, 105: 1504–1517
- Barndorff-Nielsen O E, Hansen P R, Lunde A, et al. Multivariate realised kernels: Consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. *J Econometrics*, 2011, 162: 149–169
- Wang Y, Zou J. Vast volatility matrix estimation for high-frequency financial data. *Ann Statist*, 2010, 38: 943–978
- Tao M, Wang Y, Yao Q, et al. Large volatility matrix inference via combining low-frequency and high-frequency approaches. *J Amer Statist Assoc*, 2011, 106: 1025–1040
- Tao M, Wang Y, Chen X. Fast convergence rates in estimating large volatility matrices using high-frequency financial data. *Econom Theory*, 2013, 29: 838–856
- Black F. Noise. *J Finance*, 1986, 41: 528–543
- O'Hara M. *Market Microstructure Theory*. Oxford: Blackwell, 1995
- Madhavan A. Market microstructure: A survey. *J Financ Markets*, 2000, 3: 205–258
- Stoll H. Market microstructure. In: *Handbook of Economics and Finance*. Amsterdam: Elsevier Science B. V., 2003
- Hasbrouck J. *Empirical Market Microstructure: The Institutions, Economics, and Econometrics of Securities Trading*. Oxford: Oxford University Press, 2007
- Roll R. A simple implicit measure of the effective bid-ask spread in an efficient market. *J Finance*, 1984, 39: 1127–1139
- Glosten L R, Harris L E. Estimating the components of the bid/ask spread. *J Financ Economics*, 1988, 21: 123–142
- Kijima M, Muromachi Y. Evaluation of credit risk of a portfolio with stochastic interest rate and default processes. *J Risk*, 2000, 3: 5–36
- Li Y, Xie S, Zheng X. Efficient estimation of integrated volatility incorporating trading information. *J Econometrics*, 2016, 195: 33–50
- Hansen P R, Lunde A. Realized variance and market microstructure noise. *J Bus Econom Statist*, 2006, 24: 127–161

- 27 Bickel P J, Levina E. Regularized estimation of large covariance matrices. *Ann Statist*, 2008, 36: 199–227
- 28 Bickel P J, Levina E. Covariance regularization by thresholding. *Ann Statist*, 2008, 36: 2577–2604
- 29 Keim D B, Madhavan A. The upstairs market for large-block transactions: analysis and measurement of price effects. *Rev Financ Stud*, 1996, 9: 1–36

High-dimensional integrated volatility matrix estimation for high-frequency financial data

Yan Mu, Huiling Yuan & Yong Zhou

Abstract High-dimensional integrated volatility matrix is an important statistic for portfolio allocation and financial risk management. Its estimation is a hot core issue in financial statistics and risk measurement. In this paper, we consider the integrated volatility matrix estimated problem of a large number of assets by using asynchronous high-frequency financial data modeling in microstructure noise with market information. When both the number of the assets and the average sample size of the price data on the assets go to infinity, this paper proposes new consistent estimators by using the non-overlapping intervals previous tick method and sparsity technique. Meanwhile, the new estimators have faster convergence rates, equivalently the best convergence rate without noise, than existed estimators under additive noise. The simulation compare and demonstrate that the proposed estimators perform much better than the existed volatility matrix estimators. Finally, the proposed method is applied to the index data of different Shanghai industries.

Keywords high-frequency data, high-dimensional volatility matrix, asynchronous, market microstructure, noise

MSC(2010) 60B12, 91G70

doi: 10.1360/N012016-00047