北京工業大學
BEIJING UNIVERSITY OF TECHNOLOGY

毕 业 论 文

题　　目　高频数据影响因子
　　股票价格预测

姓　　名　司徒婉儿

学　　号　21068121

指导教师　邓欣依

日　　期　2025 年 5 月 16 日

北京工业大学

毕业设计（论文）任务书

题目 高频数据影响因子股票价格预测

专业 统计学 学号 21068121 姓名 司徒婉儿

主要内容、基本要求、主要参考资料等：

一、主要内容：

根据股票价格的可预测性理论，对比传统股票价格预测模型理论，认为高频交易数据含有丰富的信息，基于更灵活的函数型数据，挖掘其影响因子助力股票价格预测。使用含有高频预测因子的混合模型对沪深 300 指数的开盘价进行短期预测，并将预测结果与传统时间序列模型结果进行比较。

二、基本要求：

（一）了解国内外研究现状，进行已有方法阐述，寻找创新点以确定论文具体研究方向。

（二）根据要求收集并整理高频交易数据。

（三）学习股票等金融知识以及所涉如时间序列、函数型数据分析等方法理论基础，为研究工作做好充足准备。

（四）在教师指导下能够独立完成论文撰写、相关理论推导、程序编写工作。

三、主要参考资料：

[1] Aït-Sahalia Y, Xiu D. Principal Component Analysis of High Frequency Data[J]. Journal of the American Statistical Association, 2019, 114(525): 287-303.

[2] 陈海强, 陈丽琼, 李迎星, 罗祥夫. 高频数据是否能改善股票价格预测?——基于函数型数据的实证研究[J]. 计量经济学报, 2021, 1(2): 427-436.

[3] 赵秀娟, 魏卓, 汪寿阳. 基于日内效应的沪深 300 股指期货套利的分析[J]. 管理科学学报, 2015, 18(1): 87-103.

完成期限：2025 年 1 月至 2025 年 6 月

指导教师签章：

专业负责人签章：

2024 年 12 月 3 日

毕业设计（论文）诚信声明书

本人郑重声明：在毕业设计（论文）工作中严格遵守学校有关规定，恪守学术规范；所提交的论文是我个人在导师指导下独立研究、撰写的成果，毕业设计（论文）中所引用他人的文字、研究成果，均已在毕业设计（论文）中加以说明；在本人的毕业生设计（论文）中未剽窃、抄袭他人的学术观点、思想和成果，未篡改实验数据。

本毕业设计（论文）和资料若有不实之处，本人愿承担一切相关责任。

学生签名：司徒婉儿 日期：2024 年 12 月 5 日

关于论文使用授权的说明

本人完全了解北京工业大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

（保密的论文在解密后应遵守此规定）

签名：司徒婉儿 导师签名：[illegible] 日期：2024 年 12 月 5 日

摘要

本文围绕股票价格预测问题，提出一种能够更有效利用高频交易数据的混合预测模型。该方法将高频交易数据序列视为函数型数据，采用样条平滑方法进行拟合，使用函数型主成分分析（FPCA）提取具有预测能力的特征因子，将其与传统自回归结构融合以构建混合预测模型。本文使用滚动窗口法进行短期预测，并针对包括平滑参数、滞后阶数及预测窗宽等关键超参数在内的不同参数组合进行测试与调优，得到最优结果。

实例分析部分基于真实股票高频交易数据展开，对比了混合模型在不同参数组合下的效果以及其与传统自回归模型在最优参数组合下的预测性能，结果表明混合模型在测试中具有更低的预测误差，体现了高频数据在提升预测精度方面的潜力。本文方法在实际数据分析中展现出良好的适用性，未来研究可能会进一步结合异常值检测、异方差建模等方法，增强模型在复杂市场环境下的稳健性与泛化能力。

关键词： 高频数据；函数型数据分析；主成分分析

Abstract

This thesis focuses on the stock price prediction problem and proposes a hybrid prediction model that can more effectively utilize high-frequency trading data. The method treats the high-frequency trading data series as functional data, fits them with spline smoothing, extracts the predictive eigenfactors using functional principal component analysis (FPCA), and integrates them with the traditional autoregressive structure in order to construct a hybrid prediction model. In this thesis, the rolling window method is used for short-term forecasting, and different combinations of parameters, including key hyperparameters such as smoothing parameters, lag orders and forecast window width, are tested and tuned to obtain optimal results.

Based on real stock high-frequency trading data, we compare the effect of the hybrid model under different parameter combinations and its prediction performance with the traditional autoregressive model under the optimal parameter combinations. The results show that the hybrid model has a lower prediction error in the test, which reflects the potential of high-frequency data to improve the prediction accuracy. The methods in this thesis show good applicability in real data analysis, and future research may further combine outlier detection, heteroskedasticity modeling, and other methods to enhance the model' s robustness and generalization ability in complex market environments.

Keywords: high-frequency data; functional data analysis; principal component analysis

目录

1. 绪论

1.1 研究背景

股票价格预测作为现代金融经济体系中投资决策与风险管理的核心环节，一直被学界和业界高度关注。股票的市场价格变动影响着资本流向、资源配置乃至宏观经济的稳定运行，因此，如何提高股票价格预测的准确性，不仅具有理论研究的价值，也具有实际投资操作上的重要意义。传统金融学理论中，有效市场假说（Efficient Market Hypothesis, EMH）[1] 占据主导地位，其基本思想在于，市场价格能够迅速且充分地反映所有可得信息，所以未来价格变动是随机、无法预测的。然而随着行为金融学和经验研究的深入，越来越多的学者开始质疑 EMH 在现实市场中的适用性[2]，尤其是在短期交易频繁，信息不对称以及投资者非理性行为较为普遍的市场环境中，股票价格呈现出可被挖掘的结构性特征和模式[3]，存在某种规律可循，从而为预测提供了可能。

传统的股价预测模型主要有自回归（AR）、滑动平均（MA）、ARIMA 等时间序列模型，这些模型依赖于日度、周度或月度等低频数据进行预测。虽然这些模型在早期研究中具有成效，但没有充分考虑到交易日内部的市场信息变动，因此无法刻画股价在更精细时间尺度上的动态特征。除此之外，随着机器学习和人工智能技术的兴起，非线性预测模型也有了更多的表现机会。神经网络、支持向量机、集成学习等预测模型突破了传统统计模型的某些限制，能够处理复杂非线性关系、高维数据、实施自动特征工程等。但传统预测方法和新兴机器学习方法都在一定程度上忽略了高频数据所蕴含的丰富信息，所以其预测性能仍然存在可提升的空间。

随着金融信息技术的快速发展及高频交易的普及，研究者们逐渐开始关注高频交易数据在股票价格预测中的应用潜力。高频交易数据指的是在极短的时间间隔内采集的，包括每分钟、每秒钟甚至每笔交易的金融市场数据，这类数据不仅包含价格、成交量等情况信息，也有反映市场的微观结构性变化的度量值，比如流动性、报价深度、订单簿行为等指标。理论研究与实证分析表明，高频交易数据在揭示市场微观机制、捕捉短期价格跳跃能力以及改进波动率建模效果等方面都具有显著优势。

在这样的背景下，函数型数据分析方法（Functional Data Analysis, FDA）被引入到高频金融数据的研究中，FDA 强调将离散观测值视为函数，从整体的函数形态出发进行分析建模，能够有效捕捉金融时间序列中的整体趋势、周期性结构以及局部异常等动态特征。在对日内价格曲线的建模中，FDA 能够避免数据的过度简化，从而保留更多信息，为建立股票价格预测模型提供更详细的依据。

1.2 研究意义

中国金融市场制度不断完善、市场交易机制变得更加复杂，在沪深 300 指数期货与 ETF 等工具日渐活跃的背景下，高频交易数据方向的研究和应用也在不断增加。一方

面，国内金融市场还未形成完善的期权定价体系，所以对于高频现货数据的研究比较重要；另一方面，量化投资的引进与发展也对高频交易数据分析方法的研发和使用提出了更高的要求。因此，如何从高维、非线性的高频数据中提取有效信息且使提取的信息具有较高质量的方法，成为了学界与业界共同的研究目标。在这样的背景下，本文从函数型数据分析出发，基于高频交易数据提取有效信息因子，结合传统时间序列预测模型构建混合模型的实际应用效果依旧稳健。

1.3 研究现状

股价预测在 20 世纪中叶已成为金融计量学中的重要议题，形成了较为系统的理论体系和实证方法。在高频交易数据普及的背景下，学者们尝试通过更精细的时间粒度和更灵活的建模框架来突破传统预测方法的精度限制，提高刻画股价短期波动的能力。

在早期的研究中，自回归模型（AR）、滑动平均模型（MA）和 ARIMA 等线性时间序列模型是主流的预测方法。这类方法假设数据具有稳定的线性结构，依赖于历史价格信息推理趋势，特点是建模简洁且易实现，但在捕捉非线性动态、波动集聚和结构突变等方面存在局限性。因此研究者尝试使用非线性模型。Neely et al.（1997）[4] 利用神经网络模型改善了外汇市场预测效果，Kim（2003）[5] 使用了支持向量机预测股指走势，两者均显示出高于传统线性模型的预测精度。同时，Tsai 和 Hsu（2010）[6] 将技术指标与分类算法相结合，提高了对涨跌趋势的判别能力。

随着机器学习方法提出与完善，金融预测领域引入了多种集成模型和深度学习架构，比如随机森林（Random Forest）、长短期记忆网络（LSTM）等。它们从海量的输入变量中自动学习数据特征并建立复杂的非线性映射，在处理高维、多源数据方面具有优势，但其在实际应用中仍然面临许多问题，比如“黑箱效应”严重、过拟合风险高、对数据特征依赖强等。

高频数据的引入则为预测模型的改进提供了新的路径。Andersen 和 Bollerslev（1998）[7] 在研究资产波动率时首次提出“实现波动率”（Realized Volatility）概念，使用高频数据计算波动率显著提高了建模精度，开启了金融计量研究对高频数据系统性利用的先河。之后，Bollerslev、Tauchen 与 Zhou（2009）[8] 进一步扩展了高频数据的应用场景，将其引入到资产风险溢价和跳跃风险建模中，研究发现高频数据中蕴含的信息能够有效揭示市场对尾部风险和极端事件的敏感程度。此外，AïtSahalia 和 Xiu（2019）[9] 提出通过主成分分析（PCA）对高频数据进行因子提取，从而提升资产定价模型的解释能力，其方法在多个实证市场中展现出较好的稳健性和普适性。

与国外研究相比，国内关于高频数据的应用起步相对较晚，但近年来也在快速发展。早期研究多聚焦于市场微观结构和交易机制，如刘勤与顾岚（2001）[10] 探讨了高频数据在刻画日内交易行为和流动性动态中的作用，指出高频数据有助于揭示交易过程中的异质性特征。随着数据可获得性的提高，研究者开始关注其在波动率与价格预测中的作用。魏宇（2010）[11] 在研究中发现，使用高频数据构建的波动率模型在预测性能上显

著优于基于日度数据的传统模型；赵秀娟等（2015）[12] 则利用沪深 300 指数期货的高频数据提取套利信号，进一步验证高频信息对市场价格反应能力的敏感性。

一些学者还尝试结合高频数据与非线性建模方法来提升预测模型的表现，例如王敏和邓华（2016）[13] 通过集成支持向量回归（SVR）与分钟级交易数据，显著提高了股指走势预测的准确率。然而大多数此类研究仍然局限于变量水平的构建，未能充分发挥高频数据作为时间函数整体的信息结构，这也为后续研究提供了方法论上的拓展空间。

函数型数据分析（Functional Data Analysis, FDA）于近年来被视为处理高频时间序列的有效工具，该方法认为观测数据在本质上具有连续的函数结构，能够通过光滑化技术捕捉日内价格曲线的整体变化趋势与细节特征。Ramsay 与 Silverman（2005）[14] 首次提出并系统地建立了 FDA 的理论框架，并将其成功应用于医学、气象等自然科学领域。Müller 等（2011）[15] 随后将 FDA 引入金融数据分析，提出将资产价格或波动率视为函数进行建模，从而克服传统时间序列方法难以处理的高维与非平稳问题。

国内学者在此基础上进行了初步探索，如陈海强等（2021）[16] 利用沪深 300 指数每分钟的高频数据，构建了结合函数型主成分分析与传统时间序列的混合预测模型，该模型在不依赖具体参数设定的前提下提取出多个日内预测因子，显著提高了对开盘价的预测精度。实证结果表明，引入函数型数据分析后的混合模型在解释力和稳定性方面均优于传统 AR 模型，验证了高频数据在提升预测性能方面的巨大潜力。

综上所述，国内外研究已在股价预测领域积累了大量成果，包括对于高频数据的利用，但是现有的研究多数侧重于提取变量层面的特征，没有充分挖掘高频数据的函数型特征，同时，如何将函数型数据结构与预测目标有效对接，构建具有解释力和实用价值的混合模型，仍然是当前研究的空白和挑战。本文基于陈海强等（2021）[16] 提出的混合预测模型，拟从函数型数据分析的角度切入，探索如何从高频数据中提取反映市场动态变化的潜在因子，解释因子特征，并结合时间序列建模思路构建股价预测模型，进行不同参数组合下的结果对比，以期在方法论和实证表现上均有所优化。

1.4 研究内容

本文主要探讨如何从高频交易数据中提取特征因子以进行股票价格预测的问题，具体而言，本文选取沪深 300 指数深市部分样本股为研究对象，将每日的高频价格曲线视作一个函数型数据，通过样条平滑后，使用函数型主成分分析（FPCA）提取日内走势中的主要变异方向，构建兼具稳定性与解释力的预测因子。同时，引入传统的时间序列模型，构建函数型数据与时间序列结构并行融合的混合预测模型，以预测下一交易日的开盘价。

在实证层面，本文通过滚动窗口预测实验，评估所建模型在多个时期内的预测表现，且通过不同参数组合测试调优，获得最优精度，最终与传统自回归模型进行预测性能对比，通过误差指标（MARE 和 MSRE）的系统评估，验证高频数据函数型因子在实际预测中所带来的增量价值。

1.5 文章结构

本文围绕构建含有高频数据因子的股票价格预测混合模型展开。第一章阐述了本课题的研究背景与意义、国内外研究现状等，随后给出了本文研究内容和文章结构，有助于读者掌握本文关键信息概要。第二章介绍了构建模型所需要的预备知识，主要介绍了函数型主成分分析理论，样条平滑方法等技术，给出函数型主成分分析理论的定义以及因子的提取方法、模型预测方法以及性能衡量指标等，为后续模型推理提供基础。在第三章中我们展示了模型构建过程及参数估计方法。第四章进行实证分析，在真实数据集上应用混合模型，进行数据处理、因子提取等操作后，通过不同参数组合的结果对比选出了最优参数组合，并且将其与传统自回归模型预测结果进行对比。在结论部分，我们给出了本文工作的全面总结。

2. 预备知识

2.1 函数型数据与样条平滑

2.1.1 函数型数据的概念

传统数据分析多处理标量或向量形式的数据，如每日开盘价、收益率等。然而在高频金融情境中，如果将每日交易价格轨迹看作一个完整的曲线或“函数”，则每个观测对象不再是单一数值，而是一条在时间区间上变化的函数曲线。

形式上，若第 i 日的价格变化为函数 $X_i(t)$，则整个样本可表示为 $\{X_1(t), X_2(t), \ldots, X_n(t)\}$。这类数据通常在固定时间点上离散观测，如每分钟 1 个点，四个小时共 240 个点。通过样条平滑等技术，可将这些离散点还原为近似连续的函数表示，从而便于整体建模。

函数型数据分析的优势包括其考虑整体特征而不仅依赖于少数固定时间点的测量值；便于利用函数的性质进行降维，处理自变量和因变量之间复杂的线性关系等；也便于通过可视化增强对于数据的理解和解释。

2.1.2 样条函数的基本原理

现实金融市场中，高频交易数据的采样频率非常高，容易受到微观市场机制、信息噪声、交易跳跃等因素的干扰，造成价格曲线呈现剧烈波动的现象，直接使用原始点值数据建模可能引发模型的不稳定与过拟合问题，因此需要在保持原数据基本趋势的同时进行适当的平滑处理。

样条函数是一种分段拟合技术，基本思想是在不同子区间内采用低阶多项式进行拟合，通过连接节点（“结点”）来实现整体的连续性和平滑性。最常见的是 B 样条（B-spline）和自然样条（natural spline），前者具有良好的局部拟合特性，后者在区间两端引入边界约束，减少边缘处拟合的不稳定。

设数据集为 $(x_i, y_i),\ i = 1, \ldots, n$，样条函数可表示为：$f(x) = \sum_{j=1}^{K} \beta_j B_j(x)$, 其中，$B_j(x)$ 为第 j 个基函数，β_j 为对应的系数，K 为基函数个数。

2.1.3 平滑样条与惩罚方法

除分段拟合外，更进一步的做法是引入平滑惩罚项，在损失函数中加入函数曲率的惩罚，以控制拟合曲线的“粗糙程度”。其常见形式如下：

$$\min_{f \in \mathcal{H}} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int (f''(x))^2 dx,$$

上述表达中，第一项为最小化拟合误差，第二项则限制函数的二阶导数，即惩罚函数的弯曲程度。λ 是平滑参数，其值越大，函数越光滑但可能欠拟合；反之，若其趋近于 0，则拟合函数可能过度追踪数据中的局部噪声。

2.2 函数型主成分分析（FPCA）

2.2.1 FPCA 的动机与定义

由于函数型数据为无限维对象，其分析计算面临维度灾难问题。为此，函数型主成分分析（FPCA）应运而生。FPCA 是 PCA（主成分分析）在函数空间的自然延伸，其目标是通过寻找最优的正交函数组合，对原始函数数据进行压缩和特征提取。

设函数型数据 $X_i(t)$ 的均值函数为 $\mu(t)$，协方差函数为 $C(s,t)$，则 FPCA 的基本展开形式为：

$$X_i(t) = \mu(t) + \sum_{k=1}^{K} \xi_{ik}\phi_k(t),$$

其中，$\phi_k(t)$ 为对应的特征函数，ξ_{ik} 为载荷（Loading），反映第 i 个样本在第 k 个主成分方向上的变化强度。

2.2.2 主成分个数选择与解释

在函数型主成分分析中，首先计算样本的协方差函数，刻画不同时间点之间的协变关系。其次，通过求解协方差函数的特征值，获得一组正交特征函数。这组特征函数可以反映样本中最主要的变化模式。对于每一个预测样本，可以通过特征函数计算其在各主成分方向上的得分，得分数值表明其在对应变化模式上的表现强度。

我们常使用均值函数及若干主成分函数与得分加权和来近似原始曲线以对函数型数据进行有效近似重构，主成分的选取遵循累计方差解释率（CRV）标准，定义为前 K 个主成分解释的总方差与整体总方差的比值，即累计贡献率。我们以累计贡献率超过 90% 为标准确定主成分的个数，这种方式在保持数据主要特征的同时又避免了模型结构的过度复杂化。

此外，主成分在经济含义上也有明确的解释，例如第一主成分通常反映股价整体的涨跌趋势，是价格变动中最为显著的共性因素；第二主成分往往揭示日内不同交易时段（如早盘与尾盘）之间的波动差异。

2.2.3 滚动预测方法

滚动窗口预测是一种常用于金融时间序列预测的实用方法。设定训练窗口长度为 w，在时间序列上逐步向前滑动进行建模与预测。较大的窗宽包含更多历史信息，参数估计更稳定，凡是对市场结构的变化的响应趋于迟缓。较小的窗宽能够快速捕捉市场机

制变化，同时也可能受噪声干扰较大。基本步骤如下：

1. 用第 t 到第 $t+w-1$ 个观测构建模型；

2. 用第 $t+w$ 个观测进行一次步预测；

3. 将训练窗口前移一步，重复上述过程.

2.2.4 预测误差指标

本文采用平均绝对相对误差 (Mean Absolute Relative Error, MARE) 和均方相对误差 (Mean Square Relative Error, MSRE) 评估模型预测的准确性，前者衡量预测值与真实值之间的相对误差的平均大小，能直观展示预测结果在整体上偏离真实值的平均幅度，对于极端值不太敏感；后者则使较大的误差被更显著地放大，适用于对预测精度有较大需求的情形。

- 平均绝对相对误差（MARE）：

$$\mathrm{MARE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{\hat{y}_t - y_t}{y_t}\right|.$$

- 均方相对误差（MSRE）：

$$\mathrm{MSRE} = \frac{1}{n}\sum_{t=1}^{n}\left(\frac{\hat{y}_t - y_t}{y_t}\right)^2.$$

通过对不同平滑参数、预测窗宽、滞后阶数、基函数个数的参数组合进行滚动实验，本文评估函数型建模在股价预测中的有效性与稳健性。

2.3 本章小结

本章我们主要介绍了后续建立模型过程中会运用的一些理论知识，包括函数型数据与样条平滑方法，函数型主成分分析理论，预测方法及指标介绍。其中第一部分主要介绍了函数型数据的概念、数学表达形式及转化方法；第二部分主要介绍了函数型数据主成分分析的思想与操作，这是本文提取高频数据因子的重要部分；第三部分主要介绍了滚动预测的原理及预测误差指标的选择，最后给出了试验设计方法。为后文模型介绍及估计方法的给出奠定基础，同时方便读者理解。

3. 模型介绍和模型估计

3.1 模型介绍

本文使用的结合高频数据因子与传统自回归模型的混合模型[16] 通过函数型主成分分析从分时交易数据中提取日内预测因子，将预测因子与自回归模型（AR）相结合从而对股票开盘价格进行短期预测。具体模型如下：

$$Y_i = \beta_0 + \sum_{l=1}^{p} \beta_l Y_{i-l} + \int_0^T (X_{i-1}(t) - M_{i-1}) g(t) dt + \epsilon_i, \tag{1}$$

其中，Y_i 是第 i 天开盘价，$X_{i-1}(t)$ 表示第 $i-1$ 天内第 t 分钟的高频股票价格，M_{i-1} 是第 $i-1$ 天的平均交易价格，其定义为日内分钟交易价格的均值，日内分钟交易价格定义为每分钟交易价格最高与最低价之均值。$\beta_0, \beta_1, \ldots, \beta_p$ 是需要估计的系数，$g(t)$ 是一个未知响应函数，ϵ_i 是残差项。在构建模型的过程中，我们需要对系数 β 序列和未知函数 $g(t)$ 进行参数估计以形成实际的预测模型。从公式（1）可知，该混合模型前半部分为 AR(p) 时间序列模型，后半部分为包含残差项的函数型数据模型。

我们把最后两项的和定义为 e_i，模型（1）可以简略地表示为：

$$Y_i = \beta_0 + \sum_{l=1}^{p} \beta_l Y_{i-l} + e_i, \tag{2}$$

其中，

$$e_i = \int_0^T (X_{i-1}(t) - M_{i-1}) g(t) dt + \epsilon_i. \tag{3}$$

注意到 $X_{i-1}(t)$ 表示第 $i-1$ 天 t 分钟的高频股票价格，M_{i-1} 是第 $i-1$ 天的平均价格，去除总体均值影响可以保证 $\mathbb{E}[\int_0^T (X_{i-1}(t) - M_{i-1}) g(t) dt] = 0$，从而满足 $\mathbb{E}[e_i] = 0$，这个条件，保证了模型的可识别性（identifiability）。

3.2 模型估计

我们定义 $Z_i(t) = X_{i-1}(t) - M_{i-1}$，先对 $Z_{i-1}(t)$ 进行光滑化处理，再对光滑曲线 $Z_{i-1}(t)$ 展开，有：

$$Z_i(t) = \sum_{k=1}^{\infty} \xi_{ik} \phi_k(t), \tag{4}$$

其中，$\phi_k(t)$ 是特征函数，满足以下正交条件：

$$\int \phi_k(t)\phi_k(s)dtds = 1; \int \phi_k(t)\phi_m(s)dtds = 0; m \neq k,$$

而 ξ_{ik} 是载荷（loading），可以通过：

$$\xi_{ik} = \int \phi_k(t)Z_i(t)dt, \tag{5}$$

计算得到。此外，特征函数 $\phi_k(t)$ 与曲线 $Z_i(t)$ 的协方差函数 $\gamma(t,s)$ 存在如下关系：

$$\int \gamma(t,s)\phi_k(s)ds = \lambda_k\phi_k(t),$$

其中，λ_k 是特征函数 $\phi_k(t)$ 与曲线 $Z_i(t)$ 的协方差函数 $\gamma(t,s)$ 的第 k 个特征值，$\phi_k(t)$ 为其对应的特征函数。总体协方差函数 $\gamma(t,s)$ 是未知的，需要利用样本协方差来估计，即：

$$\hat{\gamma}(t,s) = \frac{1}{N}\sum_{i=1}^{N} Z_i(t)Z_i(s).$$

利用总体协方差函数的估计值 $\hat{\gamma}(t,s)$ 计算出 $\phi_k(t)$，同时，在实际的估计过程中，我们需要决定主成分个数 K，即用前 K 个主成分函数 $\phi_k(t), k = 1, 2, \cdots, K$ 来近似表示原函数，我们要求这 K 个主成分要能反映出原始观测变量的大部分特征，同时不过度关注细节变化，节省计算成本和储存空间，累积贡献率（cumulative contributions proportion, CCP）是常用的选择 K 的准则之一，其定义如下：

$$\text{CCP} = \frac{\sum_{k=1}^{K}\lambda_k}{\sum_{k=1}^{N-1}\lambda_k}.$$

我们取 K 为满足 $\text{CCP} > 90\%$ 不等式的最小正整数。在求得各特征向量后，根据式（5）可以计算各个体 $Z_i(t)$ 在对应主成分的得分 ξ_{ik}，即因子得分，于是可以将原函数 $Z_i(t)$ 写成：

$$Z_i(t) = \sum_{k=1}^{K}\xi_{ik}\phi_k(t). \tag{6}$$

将 $Z_i(t) = \sum_{k=1}^{K}\xi_{ik}\phi_k(t)$ 代入式（3），则有：

$$e_i = \int_0^T \left[\sum_{k=1}^{K}\xi_{(i-1)k}\phi_k(t)\right] g(t)dt + \epsilon_i = \sum_{k=1}^{K}\xi_{(i-1)k}\theta_k + \epsilon_i, \tag{7}$$

其中，$\theta_k = \int_0^T \phi_k(t)g(t)dt$, 而预测方程（1）式可以写成：

$$Y_i = \beta_0 + \sum_{l=1}^{p} \beta_l Y_{i-l} + \sum_{k=1}^{K} \xi_{(i-1)k}\theta_k + \epsilon_i. \tag{8}$$

由于 Y_{i-l} 可以观察到，而 $\xi_{(i-1)k}$ 可以通过上述函数型主成分分析方法得到，可以利用（8）式通过最小二乘法回归（OLS）估计出相应系数 $\hat{\beta}_l$ 以及 $\hat{\theta}_k$。

接下来估计系数函数 $g(t)$，我们将 $g(t)$ 用样条基函数展开为：

$$g(t) = \sum_{m=1}^{M} b_m B_m(t), \tag{9}$$

其中 $B_m(t)$ 为样条基函数。理论上可以证明样条基的选择并不重要，我们选择日内价格曲线的特征函数来作为样条基，因此有：

$$g(t) = \sum_{k=1}^{K} b_k \phi_k(t). \tag{10}$$

将上式代入 θ_k 的定义式，并利用 $\phi_k(t)$ 的正交性质，有：

$$\theta_k = \int_0^T \phi_k(t) g(t) dt = \int_0^T \phi_k(t) \sum_{k=1}^{K} b_k \phi_k(t) dt = b_k. \tag{11}$$

可以看出，θ_k 其实是 $g(t)$ 在样条基 $\phi_k(t)$ 上的投影，因此 $g(t)$ 可以近似估计为：

$$\hat{g}(t) \approx \sum_{k=1}^{K} \hat{\theta}_k \phi_k(t). \tag{12}$$

3.3 本章小结

本章较为详细地阐述了混合模型的构建过程。首先，给出了模型结构，模型分为传统时间序列自回归模型（AR）及函数型数据模型两部分，后者的期望为 0 保证了模型的可识别性。其次，根据特征函数的正交条件以及与曲线的协方差关系给出特征函数的计算方法，通过样条基展开估计出系数函数。

4. 实例分析

4.1 数据说明

沪深 300 指数由沪深证券交易所联合发布，反映了 A 股市场整体指数，指数样本选自沪深两个证券市场，覆盖了大部分流通市值，其中的成份股为市场中市场代表性好，流动性高，交易活跃的主流投资股票，能够反映市场主流投资的收益情况。

我们选取了沪深 300 指数进行数据分析，包括高频分钟交易数据与低频日度交易数据，并将结果与现有方法进行对比，下面对该数据集所包含的信息进行简要说明。

样本区间是 2023 年 1 月 3 日到 2025 年 1 月 8 日，除周末及法定节假日等休市期，共有 489 个交易日，其开盘时间为 9：30，收盘时间为 15：00，根据《深圳证券交易所交易规则》，深市在 14：57 至 15：00 为收盘集合竞价时间。在这三分钟内，交易方式为集合竞价，收盘价采用集合竞价的结果，因此不会以分钟数据的形式实时显示，而是在 15：00 统一撮合计算收盘价。故每日有 240 个交易数据，总共有 117260 条数据。

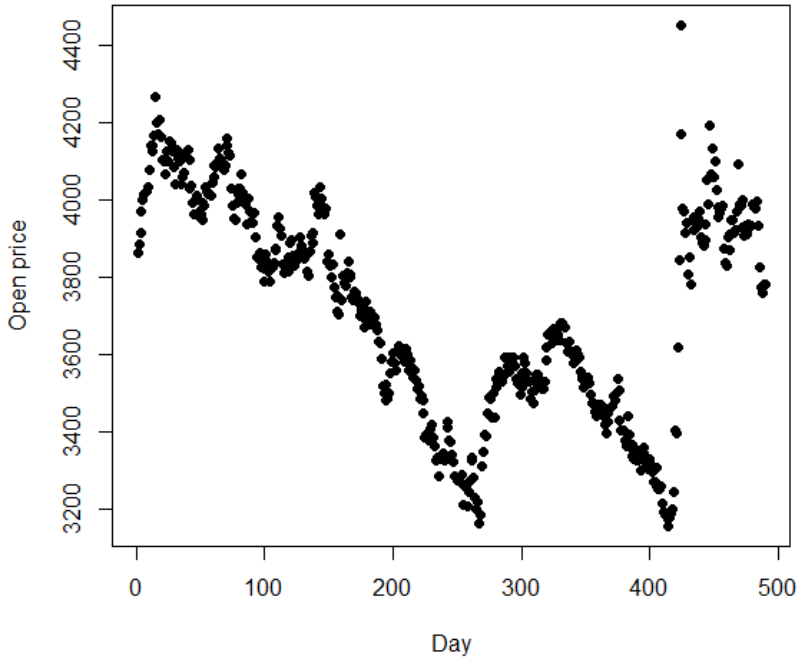


图 1: 2023 年 1 月 3 日至 2025 年 1 月 8 日沪深 300 指数深市部分样本股日开盘价

低频数据每日开盘价对应高频数据每日首分钟开盘价。图 1 画出了样本区间所有交易日的开盘价格，可以发现日开盘价在预测区间内呈现下降趋势，在第 270 及第 410 个交易日达到了最低值，随后短期内急剧上升，最后稳定在 4000 元/股价格上下。平均价格为 3708.4，最高价格为 4450.4，最低价格为 3110.0，样本区间开盘价的标准差 275.3。

4.2 数据分析

我们首先对高频数据集检查有无异常值及缺失值，绘制部分样本箱线图如下：

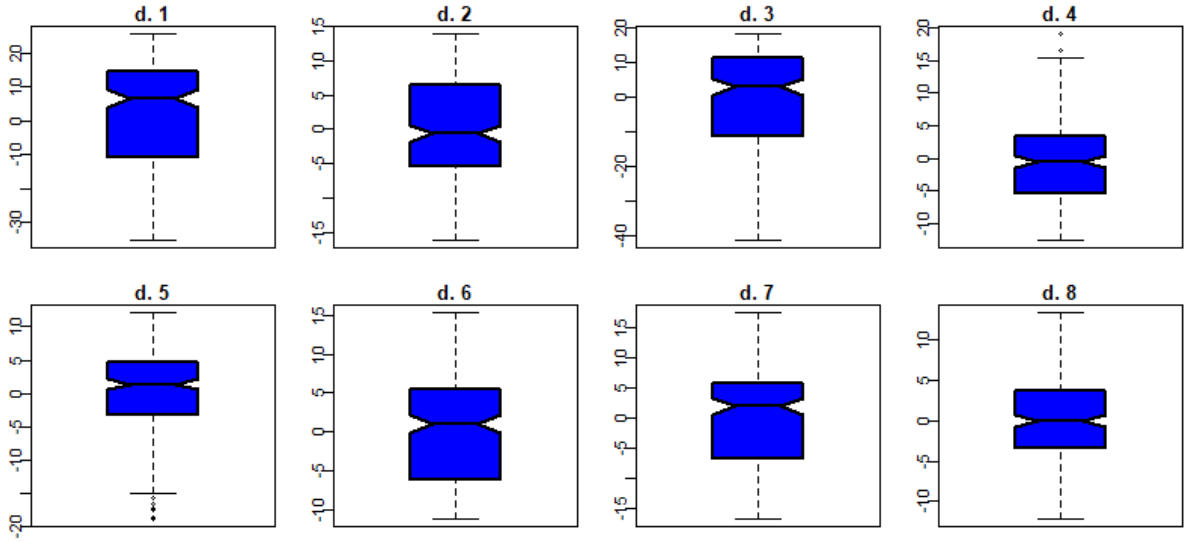
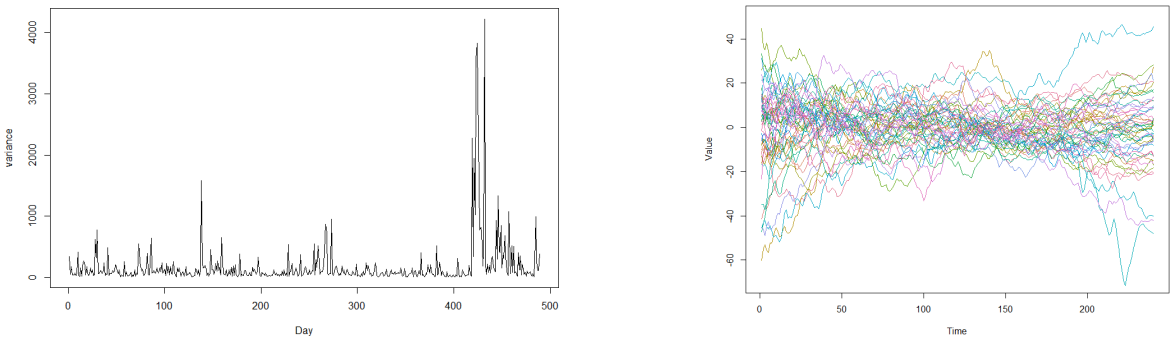


图 2: 部分数据展示

然后将数据中格式为“2023-01-03 09：30：00”的时间变量按照年份、日月、分钟拆分成 2023、0103、0930 的形式，便于后期检索。

通过计算每一交易日内价格曲线方差，绘制方差在样本区间中变化的图像。可以发现，大部分价格曲线稳定在一个较低方差水平，也存在少量日期方差在短期内急剧上升，表明市场可能在该时间点可能出现非典型事件干扰及短暂冲击，这种局部强波动现象对建模的精度和稳健性提出了更高要求，也进一步验证了滚动窗口预测方法的重要性。

此外，我们选取了样本区间前 50 个日内价格曲线进行叠加，日内价格定义如下文所述。图像显示，不同交易日的价格曲线在整体趋势、局部波动形态以及最值点分布方面存在差异，呈现出异质性，为后续分析提供了较为充分的结构依据。



(a) 每日价格曲线方差　　(b) 前 50 日价格曲线

图 3: 数据特征

我们定义每分钟的交易数据为该分钟内最高价格与最低价格的平均，针对每日共240 分钟的交易价格，把一天的分时交易价格数据看作独立重复实验的观测值，将每天的分时交易价格减去当日价格的均值，并利用样条平滑方法对其进行平滑。我们选取了6（0.001，0.01，0.1，1，10，100）个不同的平滑参数进行效果对比。如下是其中 6 个交易日平滑在平滑参数为 0.01 时的平滑日内分钟交易价格曲线。可以看到曲线在平滑处理的同时较好地保留了局部特征，便于后续预测的特征捕捉。

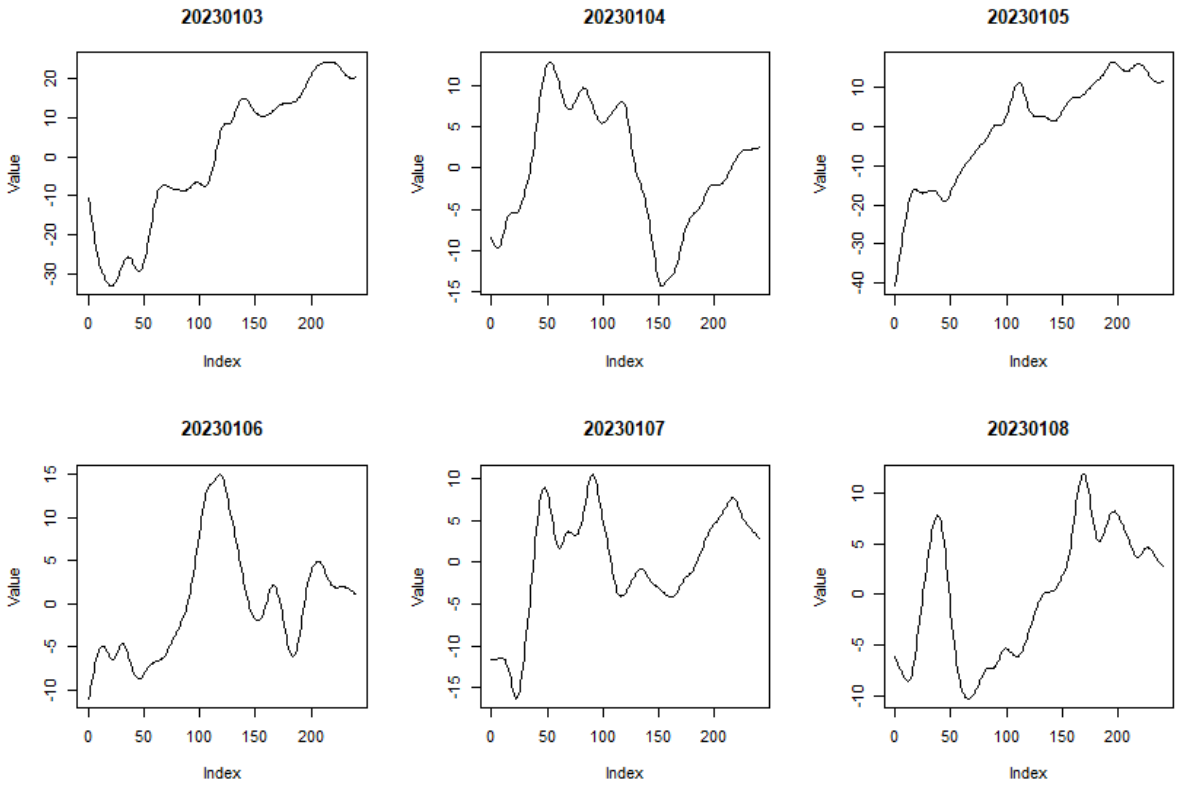


图 4: 日内交易价格曲线

接下来，对于每日的分时价格曲线，我们都以达到 90% 累积贡献率的标准选择主成分个数，提取对应的日内因子得分，作为变量输入后续的混合模型。实验中发现，该数据集数据经不同参数组合进行样条平滑与主成分提取后皆在第四个主成分达到 90% 累计贡献率。

下图画出了第一日的前四个主成分所对应的特征函数曲线图。可以看到，左上角的图是第一个主成分所对应的特征函数曲线，整体呈现下降趋势，且开市部分有短暂的比较陡峭的下降，第一个特征函数的贡献率达到 63.46%，说明一日内趋势是最重要的预测变量。如果因子载荷的系数为负，该因子也可以表示出上涨趋势。第二幅图对应是 V 型翻转的特征函数，即开市之后股票价格短期内快速下跌后又迅速上涨的数学形态，转折点在一日的正中左右，该因子解释大约 15.19% 的变化。随后两个主成分的特征函数图像是周期波动因子，但对应不同的频率，分别解释 9.3%、5.82% 的变化。四个主成分的累计贡献率超过 90%，可以较好地解释第一日的价格变动特点。

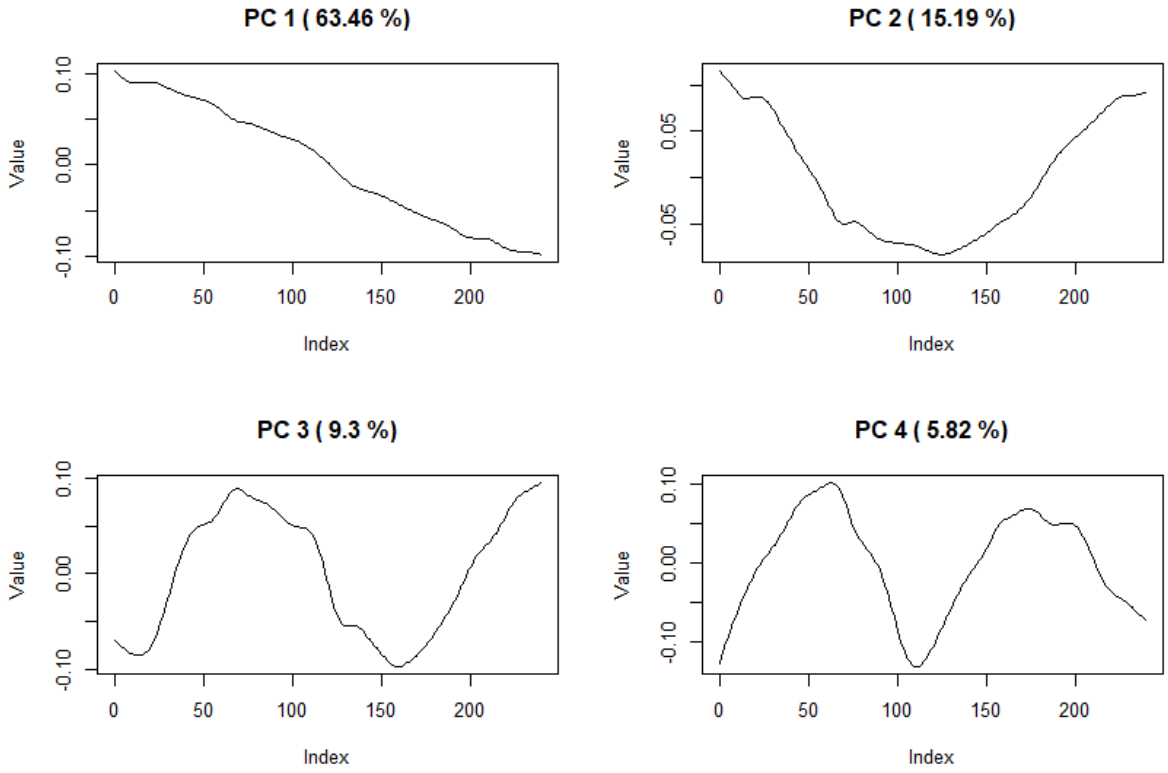


图 5: 特征向量曲线

我们将所获得的日内因子加入混合模型，使用滚动窗口法 (rolling) 进行预测，我们选取由 150 至 300 间隔为 50 的四个窗宽，用于效果对比。具体而言，以窗宽 200 为例，样本区间一共 489 个交易日，用前 200 个交易日的数据来预测下一个交易日的开盘价。比如，利用第 1 个交易日到第 200 个交易日的数据来预测第 201 个交易日的开盘价，依次类推，可以得到对 2024 年的 288 个交易日开盘价的预测值，即第 201 到第 489 个交易日的开盘价预测值。我们将这 288 个预测数据与对应的真实开盘价相比较，并与传统 AR(p) 模型的预测进行对比，来评估本文涉及的混合模型是否具有更好的预测表现，从而回答高频数据是否有助于改善股价预测这一问题。

4.3 结果分析

4.3.1 窗口长度对预测效果的影响分析

在本节中，我们对于不同窗宽预测的结果进行分析。本文报告的连续变量均采用两位小数精度，实验发现，基函数的个数（15，25，35）对于预测结果的影响不大。因此固定基函数个数为 25，对于不同的平滑参数及滞后阶数进行结果对比，结果如表所示。

表 1: 窗宽 150，基函数个数 25，不同平滑参数及滞后阶数的结果对比

lambda	MARE				MSRE			
	p=1	p=2	p=3	p=4	p=1	p=2	p=3	p=4
0.01	0.55%	0.58%	0.60%	0.61%	0.09‰	0.11‰	0.13‰	0.14‰
0.1	0.55%	0.58%	0.60%	0.60%	0.09‰	0.11‰	0.13‰	0.13‰
1	0.55%	0.56%	0.58%	0.58%	0.09‰	0.10‰	0.12‰	0.12‰
10	0.77%	0.77%	0.77%	0.77%	0.13‰	0.13‰	0.13‰	0.13‰

表 2: 窗宽 200，基函数个数 25，不同平滑参数及滞后阶数的结果对比

lambda	MARE				MSRE			
	p=1	p=2	p=3	p=4	p=1	p=2	p=3	p=4
0.01	0.52%	0.55%	0.56%	0.57%	0.09‰	0.11‰	0.12‰	0.12‰
0.1	0.52%	0.55%	0.56%	0.56%	0.09‰	0.10‰	0.12‰	0.12‰
1	0.52%	0.53%	0.54%	0.54%	0.09‰	0.10‰	0.11‰	0.11‰
10	0.73%	0.73%	0.73%	0.73%	0.13‰	0.13‰	0.13‰	0.13‰

表 3: 窗宽 250，基函数个数 25，不同平滑参数及滞后阶数的结果对比

lambda	MARE				MSRE			
	p=1	p=2	p=3	p=4	p=1	p=2	p=3	p=4
0.01	0.55%	0.58%	0.59%	0.59%	0.10‰	0.12‰	0.13‰	0.13‰
0.1	0.55%	0.57%	0.59%	0.59%	0.10‰	0.11‰	0.13‰	0.13‰
1	0.54%	0.55%	0.56%	0.56%	0.10‰	0.10‰	0.12‰	0.12‰
10	0.73%	0.73%	0.73%	0.73%	0.15‰	0.15‰	0.15‰	0.15‰

表 4: 窗宽 300，基函数个数 25，不同平滑参数及滞后阶数的结果对比

lambda	MARE				MSRE			
	p=1	p=2	p=3	p=4	p=1	p=2	p=3	p=4
0.01	0.56%	0.59%	0.60%	0.61%	0.11‰	0.12‰	0.13‰	0.14‰
0.1	0.56%	0.58%	0.60%	0.60%	0.11‰	0.12‰	0.13‰	0.13‰
1	0.55%	0.55%	0.57%	0.57%	0.11‰	0.11‰	0.12‰	0.12‰
10	0.73%	0.74%	0.74%	0.74%	0.15‰	0.16‰	0.16‰	0.16‰

当窗宽为 150，平滑参数为 0.01 时，不同滞后阶数的平均绝对相对误差 MARE 结果差异比平滑参数为 1 时的大，而 0.01 和 0.1 情形下的 MARE 基本相同，但是在滞后阶数为 4 时还是略有下降。MSRE 的数值极小，以千分率作为单位，可以看出随着平滑参数提高，均方相对误差也在减小，平滑参数为 1 时效果是比较好的。

当窗宽为 200，平滑参数选择 1 在四个滞后阶数情形下 MARE 都显著较小，且平滑参数为 10 时，虽然误差仍然较大，但是对比窗宽为 150 时效果明显更佳。MSRE 在平滑参数较小情形下显著减少，且滞后阶数的增大对其影响变小。

窗宽为 250 时，各参数组合的 MARE 都有不同程度的增大，除了平滑参数为 10 时的组合，对于特征的捕捉不明显，有同样的误差程度，平滑参数为 1 时仍然是误差较小的情形。MSRE 则在平滑参数较大时出现了对比较窗宽 200 时显著的上涨，其余参数组合分别出现 0.01‰ 的数值上涨。

当窗宽增大到 300 时，误差同样出现了较明显的上涨。因此可以认为在股票价格短期预测这种高频、即时的场景下，滞后阶数越小越好。

不论窗宽的取值如何，平滑参数为 10 时的 MARE 显著比别的情形误差要大，且滞后阶数的选择对其并无影响，窗宽 200 以上会比窗宽 150 时误差更小，可能是因为较短的窗宽更容易捕捉短期波动特征，而平滑参数越大，函数越平滑，特征保留不足，因此对于结果的预测更不准确。在各参数组合下，两种误差衡量指标基本都随着滞后阶数的增大而增长。

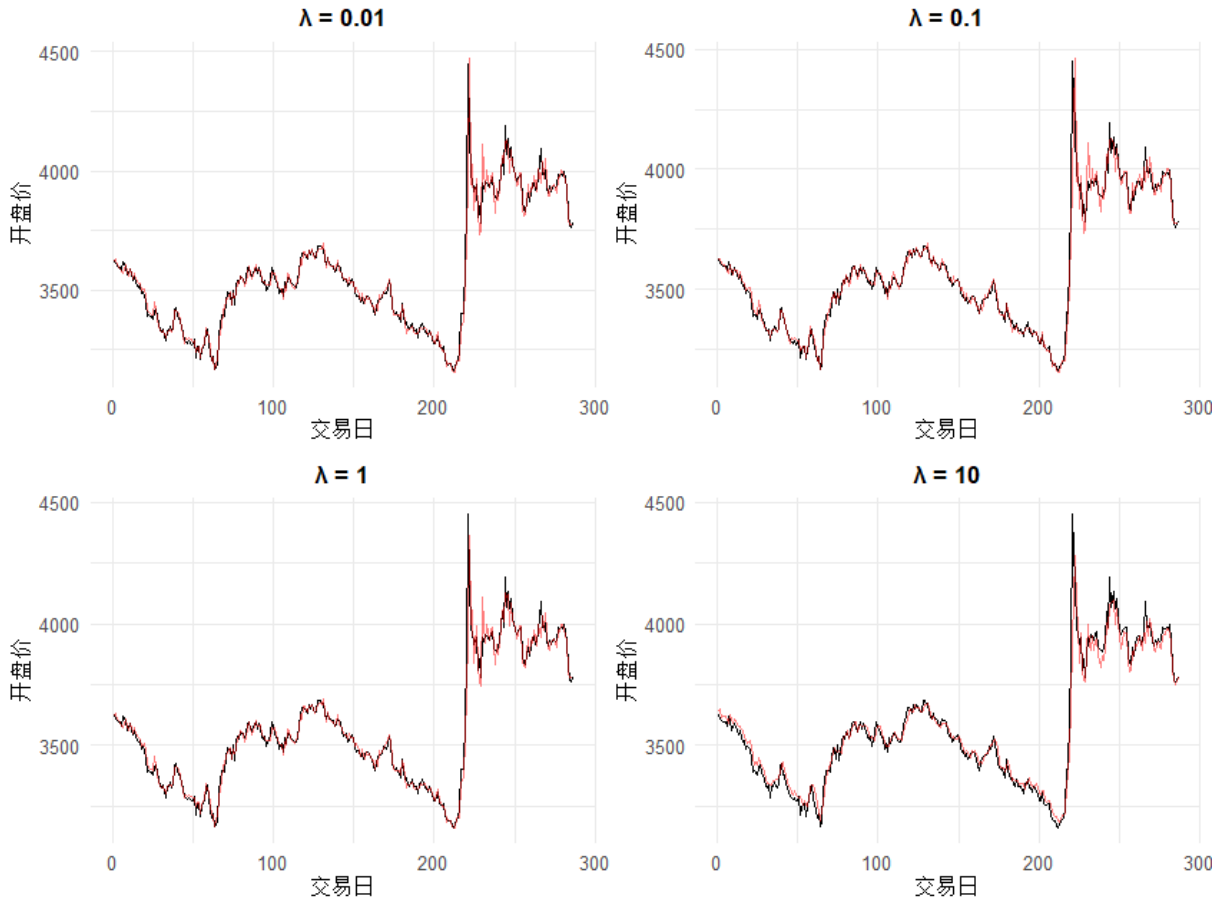


图 6: 混合模型预测与真实值对比图（p=2, 窗宽 =200）

因此我们可以认为，窗宽为 200，平滑参数为 1 时的参数组合是比较好的。

我们对窗宽 200，滞后阶数为 2 的混合模型的预测效果进行评估，上图为不同平滑

参数下混合模型对开盘价的预测值和真实值的对比，因为混合模型的自回归部分的阶数是 p=2，因此记为 $AR(2)\_F$。图中黑线是第 201 天到第 489 天的真实开盘价，红线是混合模型 $AR(2)\_F$ 对第 201 天到 489 天的开盘价预测值，可以看出红线与黑线重合度较高。由表 1 可知，前三个平滑参数的混合模型 $AR(2)\_F$ 的预测值与真实值的平均相对误差（MARE）为 0.53% 左右，说明混合模型的预测结果很接近真实值。

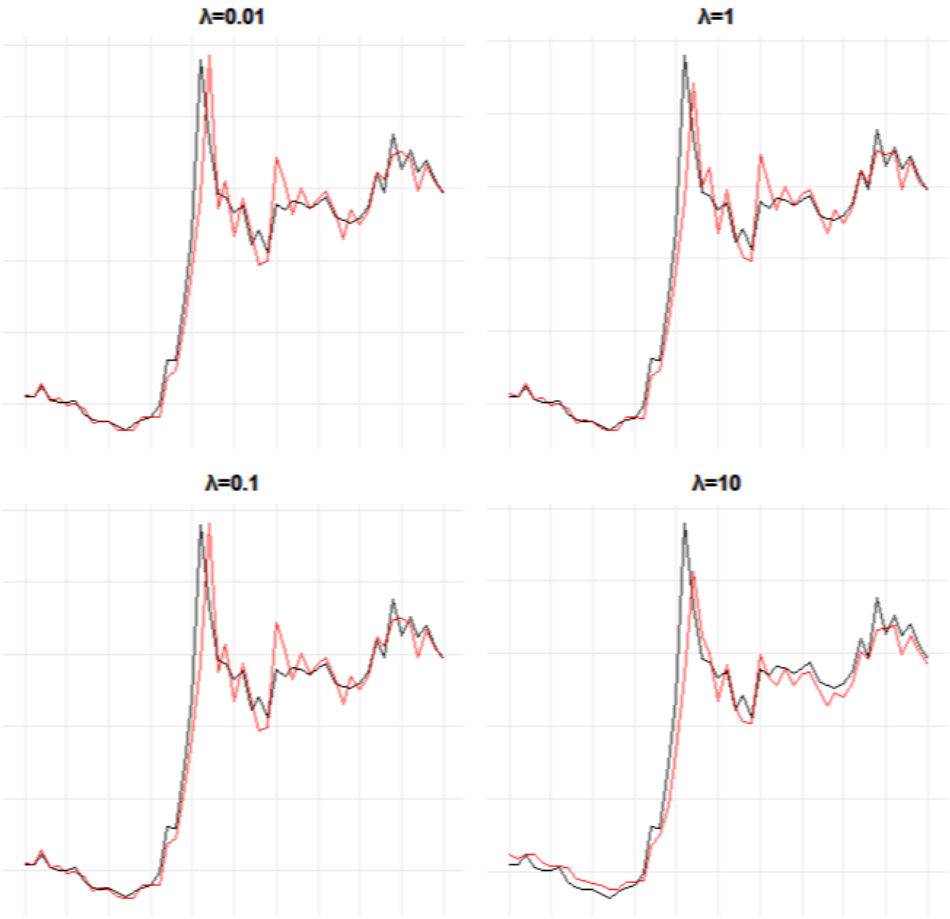


图 7: 混合模型预测与真实值局部对比图（p=2, 窗宽 =200）

根据局部细节表现，我们也可以发现不同平滑参数下的预测效果存在细微不同，比如最高峰的估计存在滞后高估、滞后、滞后低估的区别。

4.3.2 混合模型预测性能对比

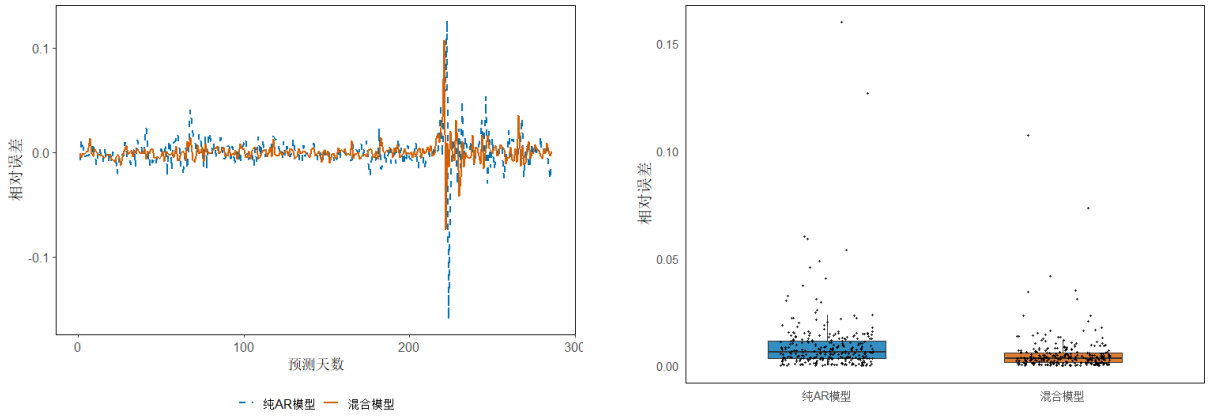
根据每日开盘价 $y_i, i = 1, 2, \cdots, 489$ 和日内交易价格曲线，我们可以估计出模型的参数。表 5 给出了窗宽 200 时两种模型在预测第 201 个交易日至第 489 个交易日的开盘价格时的平均绝对相对误差和均方相对误差。其中，$AR(p)$ 表示自回归模型，$AR(p)\_F$ 表示我们的函数型自回归模型，p 为自回归的阶数。可以看出，对所有 p 的取值，$AR(p)\_F$ 模型的预测结果较之 $AR(p)$ 模型都有很大的改进，说明高频数据得到的预测因子能够改善股价的预测能力。

表 5: 各模型预测误差比较

模型		p=1	p=2	p=3	p=4
MARE	AR(p)	0.91%	0.96%	0.98%	0.99%
	AR(p)_F	0.52%	0.54%	0.56%	0.57%
MSRE	AR(p)	0.23‰	0.29‰	0.34‰	0.35‰
	AR(p)_F	0.09‰	0.10‰	0.12‰	0.12‰

下图为我们的混合模型 $AR(2)\_F$ 和传统的时间序列模型 $AR(2)$ 在整个预测周期内的相对预测误差比较，相对误差计算方法为:（真实值-预测值）/真实值。蓝色虚线为 $AR(2)$ 的相对预测误差，红色实线为混合模型 $AR(2)\_F$ 的相对预测误差。相对误差在 0 的上下波动，两者在一定程度上具有相似的起伏，这可能是价格走势突变与稳定模型之间结构性的差别。同时，混合模型的相对误差波动更小，更趋近于 0，表明其预测值更接近真实值，预测效果更好。

通过误差的箱线图同样可以发现这样的特征，纯自回归模型的预测误差由更多离群值，且箱体更长，分布更广，混合模型的预测误差则更集聚，且数值更小。



(a) 折线图　　(b) 箱线图

图 8: 相对误差对比

4.4 本章小结

本章我们在实际数据集上应用了提出的混合模型，展示了数据说明、处理、因子提取、滚动预测以及结果分析全过程。在结果分析中，通过预测误差数据表格和预测值与真实值曲线对比了不同参数组合的模型预测效果，给出了较优的参数组合。使用该参数组合，由折线图及箱线图的对比展示了与传统自回归模型结果之间的异同，从不同滞后阶数下混合模型皆优于传统自回归模型，得出了混合模型具有更好的表现的结论。

结论

本文针对股票价格预测，给出了高频交易情形下能够更充分利用高频交易数据的预测方法，建立含有高频数据价格预测因子的混合预测模型，给出推导过程以及参数估计方法，并将模型应用于真实数据集。具体方法是对高频交易数据进行样条平滑，将其视作函数型数据并作主成分分析，提取日内预测因子，运用滚动窗口法进行短期预测。

在建模过程中，有平滑参数、滞后阶数、预测窗宽等可选参数，通过测试不同参数组合下的模型预测效果，选出了较好的参数组合。最后将含有高频数据银子的混合模型与传统自回归模型进行结果比较，验证了方法的有效性。

在真实的股票交易市场，高频数据逐渐普及，对于更加丰富且精准的价格预测方法的需求也在不断增强，本文通过理论及实践展现了其中一种有效方法，然而现实数据处理中仍会面对很多异常波动，需要结合更多方法以提高预测的准确性和稳健性，这将是我进一步要研究的问题。

参考文献

[1] Fama E F. Efficient capital markets: A review of theory and empirical work[J]. The Journal of Finance, 1970, 25(2): 383-417.

[2] De Bondt W F M, Thaler R H. Does the stock market overreact?[J]. The Journal of Finance, 1985, 40(3): 793-805.

[3] Jegadeesh N, Titman S. Returns to buying winners and selling losers: Implications for stock market efficiency[J]. The Journal of Finance, 1993, 48(1): 65-91.

[4] Neely C J, Weller P A, Dittmar R. Is technical analysis in the foreign exchange market profitable? A genetic programming approach[J]. Journal of Financial and Quantitative Analysis, 1997, 32(4): 405-426.

[5] Kim K J. Financial time series forecasting using support vector machines[J]. Neurocomputing, 2003, 55(1-2): 307-319.

[6] Tsai C F, Hsiao Y C. Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches[J]. Decision Support Systems, 2010, 50(1): 258-269.

[7] Andersen T G, Bollerslev T. Answering the skeptics: Yes, standard volatility models do provide accurate forecasts[J]. International Economic Review, 1998, 39(4): 885-905.

[8] Bollerslev T, Tauchen G, Zhou H. Expected stock returns and variance risk premia[J]. The Review of Financial Studies, 2009, 22(11): 4463-4492.

[9] Aït-Sahalia Y, Xiu D. Principal component analysis of high-frequency data[J]. Journal of the American Statistical Association, 2019, 114(525): 287-303.

[10] 刘勤, 顾岚. 高频数据与我国股票市场微观结构研究 [J]. 金融研究, 2001(11): 96-104.

[11] 魏宇. 高频数据在波动率建模中的应用研究 [J]. 数量经济技术经济研究, 2010(4): 68-75.

[12] 赵秀娟, 王永中, 许庆瑞. 高频数据在套利策略构建中的应用——基于沪深 300 指数期货的实证分析 [J]. 金融研究, 2015(2): 135-150.

[13] 王敏, 邓华. 基于高频数据的支持向量回归在股市预测中的应用 [J]. 系统工程理论与实践, 2016, 36(1): 85-92.

[14] Ramsay J O, Silverman B W. Functional Data Analysis (2nd ed.)[M]. New York: Springer, 2005.

[15] Müller H G, Stadtmüller U, Yao F. Functional data analysis for sparse longitudinal data[J]. Journal of the American Statistical Association, 2011, 96(454): 577-590.

[16] 陈海强, 陈丽琼, 李迎星, 罗祥夫. 高频数据是否能改善股票价格预测？——基于函数型数据的实证研究 [J]. 计量经济学报, 2021, 1(2): 427-435.

致谢

本论文的完成离不开许多人的支持与帮助。在此，我谨向所有在我本科期间给予指导和关心的家人、老师和朋友们表示由衷的感谢。

首先，我要感谢我的导师邓欣依老师，您在课题选题、研究思路以及论文撰写的各个阶段给予了耐心指导和宝贵建议，使我能够顺利完成本次研究。同时，我也要感谢暑期实习期间的带教老师，是您的培养和训练让我在文献阅读、汇报、学术思考等方面的能力有所提升。

其次，感谢学院提供的学习与研究平台、实验数据和技术资源，为本论文的实证分析提供了有力支持。也感谢我的同学和朋友们在研究过程中给予的交流与鼓励，你们的建议和陪伴让我受益匪浅。

最后，感谢一直站在我身后的家人，你们的支持与鼓励给予了我前进的的动力。感谢眼里有光的自己，愿你把握现在，勤学善思，博学笃行，一直创造，不断求索，勇于突破。

致以诚挚的谢意！