VOL. 4, NO. 8

OMI Research Newsletter – May 2025

The Oxford-Man Institute of Quantitative Finance (OMI) is a world-leading centre for interdisciplinary research in financial markets. The OMI addresses fundamental problems in quantitative finance with a strong focus on machine learning and data driven models.

*The **OMI Research Newsletter** is a curated summary that covers the most recent research, insights, and tools related to financial decision making.*

Announcements

The Machine Learning in Quantitative Finance Conference 2025, organised by OMI, will be held on 5–6 June 2025 at the Mathematical Institute, University of Oxford. Please find the schedule and register here.

The Workshop on Big Data and Finance will be held at OMI on 16–17 October 2025.

Highlights

Key contributions in this month's edition: Li et al. (2025) develop a theoretically sound and practically reliable volatility estimator based on intraday candlestick charts. Aït-Sahalia, Jacod and Xiu (2025) extend the classical Fama-MacBeth approach to continuous-time asset pricing models. Olea et al. (2025) provide a comprehensive comparison and practical guidance on two cornerstone tools in macroeconometrics—local projections and VARs. Ji et al. (2025) explore potential areas where statisticians can make important contributions to the development of large language models. Zhang et al. (2025) propose an unsupervised framework that clusters tick-level limit order book data to identify distinct market participant behaviours. Mo (2025) shows that analysts improve forecasts from technological information spillovers across industries within brokerages. Cetingoz and Lehalle (2025) highlight fundamental limitations of generative models in portfolio construction, and propose a tailored pipeline that improves synthetic return generation while preserving key financial properties.

Statistics, Econometrics & Machine Learning

Realized Candlestick Wicks

Y. LI, I. NOLTE, S. NOLTE, S. YU
UNIVERSITY OF MANCHESTER, LANCASTER UNIVERSITY, OXFORD-MAN INSTITUTE - UNIVERSITY OF OXFORD

We develop a novel nonparametric estimator of integrated variance by summing up the squared wick lengths of intraday candlesticks over a fixed time interval. The proposed wick-based estimator is robust to short-lived extreme price movements, such as gradual jumps and flash crashes. We investigate the asymptotic properties of the proposed estimator, and show that its asymptotic variance is about four times smaller than the state-of-the-art differenced-return volatility (DV) estimator. We also develop a Hausman-type test for the presence of both jumps and episodic extreme price movements. Monte Carlo simulations and empirical applications further validate the practical reliability of our proposed estimator.

Continuous-Time Fama-MacBeth Regressions

Y. AÏT-SAHALIA, J. JACOD, D. XIU
PRINCETON UNIVERSITY, SORBONNE UNIVERSITÉ, UNIVERSITY OF CHICAGO

We develop an asymptotic framework for conducting inference on continuous-time asset pricing models using high-frequency returns over an increasing time horizon. Our study focuses on the identification and estimation of risk premia associated with continuous component and jumps of various size brackets. We extend the classical Fama-MacBeth regression from the discrete-time setting to a continuous-time factor model, incorporating general dynamics for factors, idiosyncratic components, and factor loadings. Our empirical analysis of U.S. equities, foreign

exchange, and commodities underscores the distinct significance of continuous and jump risk premia for the specific factors constructed within each asset class in determining expected returns.

Local Projections or VARs? A Primer for Macroeconomists

J. L. M. Olea, M. Plagborg-Møller, E. Qian, C. K. Wolf

Cornell University, Princeton University, Massachusetts Institute of Technology

What should applied macroeconomists know about local projection (LP) and vector autoregression (VAR) impulse response estimators? The two methods share the same estimand, but in finite samples lie on opposite ends of a bias-variance trade-off. While the low bias of LPs comes at a quite steep variance cost, this cost must be paid to achieve robust uncertainty assessments. VARs should thus only be used with long lag lengths, ensuring equivalence with LP. For LP estimation, we provide guidance on selection of lag length and controls, bias correction, and confidence interval construction.

Deep Learning of Point Processes for Modeling High-Frequency Data

Y. Gyotoku, I. M. Toke, N. Yoshida

University of Tokyo, Universit´e Paris-Saclay, Japan Science and Technology Agency CREST

We investigate applications of deep neural networks to a point process having an intensity with mixing covariates processes as input. Our generic model includes Cox-type models and marked point processes as well as multivariate point processes. An oracle inequality and a rate of convergence are derived for the prediction error. A simulation study shows that the marked point process can be superior to the simple multivariate model in prediction. We apply the marked ratio model to real limit order book data.

Modeling and Forecasting Realized Volatility with Multivariate Fractional Brownian Motion

M. Bibinger, J. Yu, C. Zhang

University of Würzburg, University of Macau

A multivariate fractional Brownian motion (mfBm) with component-wise Hurst exponents is used to model and forecast realized volatility. We investigate the interplay between correlation coefficients and Hurst exponents and propose a novel estimation method for all model parameters, establishing consistency and asymptotic normality of the estimators. Additionally, we develop a time-reversibility test, which is typically not rejected by real volatility data. When the data-generating process is a time-reversible mfBm, we derive optimal forecasting formulae and analyze their properties. A key insight is that an mfBm with different Hurst exponents and non-zero correlations can reduce forecasting errors compared to a one-dimensional model. Consistent with optimal forecasting theory, out-of-sample forecasts using the time-reversible mfBm show improvements over univariate fBm, particularly when the estimated Hurst exponents differ significantly. Empirical results demonstrate that mfBm-based forecasts outperform the (vector) HAR model.

Distributed Time-Varying Gaussian Regression via Kalman Filtering

N. Taddei, R. Maggioni, J. Eising, G. De Pasquale, F. Dorfler

ETH Zurich, TU Eindhoven

We consider the problem of learning time-varying functions in a distributed fashion, where agents collect local information to collaboratively achieve a shared estimate. This task is particularly relevant in control applications, whenever real-time and robust estimation of dynamic cost/reward functions in safety critical settings has to be performed. In this paper, we,adopt a finite-dimensional approximation of a Gaussian Process, corresponding to a Bayesian linear regression in an appropriate feature space, and propose a new algorithm, DistKP, to track the time-varying coefficients via a distributed Kalman filter. The proposed method works for arbitrary kernels and under weaker assumptions on the time-evolution of the function to learn compared to the literature. We validate our results using a simulation example in which a fleet of Unmanned Aerial Vehicles (UAVs) learns a dynamically changing wind field.

An Overview of Large Language Models for Statisticians

W. Ji, W. Yuan, E. Getzen, K. Cho, M. I. Jordan, S. Mei, J. E Weston, W. J. Su, J. Xu, L. Zhang

Stanford University, New York University, Meta FAIR, University of Pennsylvania, University of California, Berkeley, INRIA, Rutgers University

Large Language Models (LLMs) have emerged as transformative tools in artificial intelligence (AI), exhibiting remarkable capabilities across diverse tasks such as text generation, reasoning, and decision-making. While their success has primarily been driven by advances in computational power and deep learning architectures, emerging problems – in areas such as uncertainty quantification, decision-making, causal inference, and distribution shift – require a deeper engagement with the field of statistics. This paper explores potential areas where statisticians can make important contributions to the development of LLMs, particularly those that

aim to engender trustworthiness and transparency for human users. Thus, we focus on issues such as uncertainty quantification, interpretability, fairness, privacy, watermarking and model adaptation. We also consider possible roles for LLMs in statistical analysis. By bridging AI and statistics, we aim to foster a deeper collaboration that advances both the theoretical foundations and practical applications of LLMs, ultimately shaping their role in addressing complex societal challenges.

Finance and Economics

ClusterLOB: Enhancing Trading Strategies by Clustering Orders in Limit Order Books

Y. Zhang, M. Cucuringu, A. Y. Shestopaloff, S. Zohren

Oxford-Man Institute - University of Oxford, University of California Los Angeles, Queen Mary University of London, Memorial University of Newfoundland

In the rapidly evolving world of financial markets, understanding the dynamics of limit order book (LOB) is crucial for unraveling market microstructure and participant behavior. We introduce ClusterLOB as a method to cluster individual market events in a stream of market-by-order (MBO) data into different groups. To do so, each market event is augmented with six time-dependent features. By applying the K-means++ clustering algorithm to the resulting order features, we are then able to assign each new order to one of three distinct clusters, which we identify as directional, opportunistic, and market-making participants, each capturing unique trading behaviors. Our experimental results are performed on one year of MBO data containing small-tick, medium-tick, and large-tick stocks from NASDAQ. To validate the usefulness of our clustering, we compute order flow imbalances across each cluster within 30-minute buckets during the trading day. We treat each cluster's imbalance as a signal that provides insights into trading strategies and participants' responses to varying market conditions. To assess the effectiveness of these signals, we identify the trading strategy with the highest Sharpe ratio in the training dataset, and demonstrate that its performance in the test dataset is superior to benchmark trading strategies that do not incorporate clustering. We also evaluate trading strategies based on order flow imbalance decompositions across different market event types, including add, cancel, and trade events, to assess their robustness in various market conditions. This work establishes a robust framework for clustering market participant behavior, which helps us to better understand market microstructure, and inform the development of more effective predictive trading signals with practical applications in algorithmic trading and quantitative finance.

Cross-Industry Technological Information Spillovers in Analyst Research

H. Mo

Saïd Business School - University of Oxford

This paper provides evidence of within-brokerage spillover effects on technological information—an analyst's research on one firm benefits another analyst's research on another technologically similar firm in a different industry. To capture this effect, I develop a Technological Connectedness (TC) measure based on the similarity in patent distribution between firms covered by different analysts. I show that these spillovers improve analysts' forecast performance, especially for younger, smaller, and more innovative firms, and when analyst colleagues produce higher-quality research. I also find that analysts and their highly-connected colleagues are more likely to issue revisions contemporaneously. To establish causality, I exploit two exogenous shocks: the introduction of ChatGPT-4 and brokerage mergers. Lastly, I show that analysts' improved understanding of technological information attenuates technology momentum in stock returns.

Volatility-Managed Volatility Selling

A. Yang

Peking University

We develop volatility risk premium timing strategies that trade two assets: a volatility asset and a risk-free asset. We first analyze a benchmark portfolio that sells a constant weight of volatility assets each month. Then, we show that a volatility-managed portfolio, which reduces selling volatility assets during periods of heightened volatility, considerably enhances long-run performance. Our findings are robust across variance swaps, VIX futures, and S&P 500 straddles, and even in the presence of transaction costs. An ex-post study indicates that timing portfolios yield positive alpha and reduce exposure relative to constant-weight portfolios, mostly driven by large-volatility-spike periods rather than stable periods. Our findings help differentiate asset pricing theories on risk-return relations in the volatility asset market.

The Memorization Problem: Can We Trust LLMs' Economic Forecasts?

A. Lopez-Lira, Y. Tang, M. Zhu

University of Florida

Large language models (LLMs) cannot be trusted for economic forecasts during periods covered by their training data. We provide the first systematic evaluation

of LLMs' memorization of economic and financial data, including major economic indicators, news headlines, stock returns, and conference calls. Our findings show that LLMs can perfectly recall the exact numerical values of key economic variables from before their knowledge cutoff dates. This recall appears to be randomly distributed across different dates and data types. This selective perfect memory creates a fundamental issue—when testing forecasting capabilities before their knowledge cutoff dates, we cannot distinguish whether LLMs are forecasting or simply accessing memorized data. Explicit instructions to respect historical data boundaries fail to prevent LLMs from achieving recall-level accuracy in forecasting tasks. Further, LLMs seem exceptional at reconstructing masked entities from minimal contextual clues, suggesting that masking provides inadequate protection against motivated reasoning. Our findings raise concerns about using LLMs to forecast historical data or backtest trading strategies, as their apparent predictive success may merely reflect memorization rather than genuine economic insight. Any application where future knowledge would change LLMs' outputs can be affected by memorization. In contrast, consistent with the absence of data contamination, LLMs cannot recall data after their knowledge cutoff date.

Do Machine Learning Models Need to Be Sector Experts?

M. X. Hanauer, A. Soebhag, M. Stam, T. Hoogteijling

Technical University of Munich, Robeco, Erasmus University Rotterdam, Tilburg University

We examine heterogeneous return predictability at the industry level using machine learning models trained on a comprehensive set of firm characteristics. We compare uniform ("Generalist") models with industry-specific ("Specialist") models and introduce a "Hybrid" model that incorporates industry membership. The Hybrid model outperforms the Specialist model in out-of-sample performance, yielding higher Sharpe ratios and lower risk compared to both alternatives. Additional analyses using international data corroborate these findings. Our results indicate that the Hybrid model benefits from a better signal-to-noise ratio by combining industry awareness with broader sample sizes, improving both estimation precision and learning efficiency.

The Political Economy of Firm Networks: CEO Ideology and Global Trade

E. Kempf, M. Luo, M. Tsoutsoura

Harvard University, Bayes Business School, Washington University in St. Louis

We examine how the political ideology of corporate leaders shapes cross-border firm networks. Exploiting changes in ideological alignment between U.S. firm CEOs and foreign governments around close foreign elections, we show that U.S. firms are more likely to terminate trade relationships with countries led by governments whose ideology becomes more distant from that of their CEOs. The impact is concentrated among CEOs holding strong political views, and is particularly pronounced for shorter trade relationships, suggesting ideological alignment is more relevant in more flexible and substitutable connections. Our findings highlight the role of ideology in shaping the formation and persistence of international firm networks.

Rationalizing Dynamic Choices

H. de Oliveira, R. Lamba

Fundação Getúlio Vargas-EESP, Cornell University

An analyst observes an agent take a sequence of actions. The analyst does not have access to the agent's information and ponders whether the observed actions could be justified through a rational Bayesian model with a known utility function. We show that the observed actions cannot be justified if and only if there is a single deviation argument that leaves the agent better off, regardless of the information. The result is then extended to allow for distributions over possible action sequences. Four applications are presented: monotonicity of rationalization with risk aversion, a potential rejection of the Bayesian model with observable data, feasible outcomes in dynamic information design, and partial identification of preferences without assumptions on information.

The Social Learning Barrier

F. Brandl

University of Bonn

We consider long-lived agents who interact repeatedly in a social network. In each period, each agent learns about an unknown state by observing a private signal and her neighbors' actions in the previous period before taking an action herself. Our main result shows that the learning rate of the slowest learning agent is bounded from above independently of the number of agents, the network structure, and the agents' strategies. Applying this result to equilibrium learning with rational agents shows that the learning rate of all agents in any equilibrium is bounded under general conditions. This extends recent findings on equilibrium learning and demonstrates that the limitation stems from an inherent tradeoff between optimal action choices and information revelation rather than strategic considerations.

Mathematical and Computational Finance

Synthetic Data for Portfolios: A Throw of the Dice Will Never Abolish Chance

A. R. Cetingoz, C.-A. Lehalle

Université Paris 1 Panthéon-Sorbonne, École Polytechnique

Simulation methods have always been instrumental in finance, and data-driven methods with minimal model specification, commonly referred to as generative models, have attracted increasing attention, especially after the success of deep learning in a broad range of fields. However, the adoption of these models in financial applications has not matched the growing interest, probably due to the unique complexities and challenges of financial markets. This paper contributes to a deeper understanding of the limitations of generative models, particularly in portfolio and risk management. To this end, we begin by presenting theoretical results on the importance of initial sample size, and point out the potential pitfalls of generating far more data than originally available. We then highlight the inseparable nature of model development and the desired uses by touching on a paradox: usual generative models inherently care less about what is important for constructing portfolios (in particular the long-short ones). Based on these findings, we propose a pipeline for the generation of multivariate returns that meets conventional evaluation standards on a large universe of US equities while being compliant with stylized facts observed in asset returns and turning around the pitfalls we previously identified. Moreover, we insist on the need for more accurate evaluation methods, and suggest, through an example of mean-reversion strategies, a method designed to identify poor models for a given application based on regurgitative training, i.e. retraining the model using the data it has itself generated, which is commonly referred to in statistics as identifiability.

Learning the Spoofability of Limit Order Books With Interpretable Probabilistic Neural Networks

T. Fabre, D. Challet

Université Paris-Saclay, SUN ZU Lab

This paper investigates real-time detection of spoofing activity in limit order books, focusing on cryptocurrency centralized exchanges. We first introduce novel order flow variables based on multi-scale Hawkes processes that account both for the size and placement distance from current best prices of new limit orders. Using a Level-3 data set, we train a neural network model to predict the conditional probability distribution of mid price movements based on these features. Our empirical analysis highlights the critical role of the posting distance of limit orders in the price formation process, showing that spoofing detection models that do not take the posting distance into account are inadequate to describe the data. Next, we propose a spoofing detection framework based on the probabilistic market manipulation gain of a spoofing agent and use the previously trained neural network to compute the expected gain. Running this algorithm on all submitted limit orders in the period 2024-12-04 to 2024-12-07, we find that 31% of large orders could spoof the market. Because of its simple neuronal architecture, our model can be run in real time. This work contributes to enhancing market integrity by providing a robust tool for monitoring and mitigating spoofing in both cryptocurrency exchanges and traditional financial markets.

Acknowledgement

The OMI Research Newsletter is curated by Álvaro Cartea, Director of the Oxford-Man Institute, and Shifan Yu, postdoctoral researcher at the institute. Paper suggestions are from fellow members at the institute: Álvaro Arroyo, Patrick Chang, Fayçal Drissi, Gerardo Durán-Martín, and Qi Jin.

References

Aït-Sahalia, Yacine, Jean Jacod, and Dacheng Xiu. 2025. "Continuous-Time Fama-MacBeth Regressions." *Review of Financial Studies*, forthcoming.

Bibinger, Markus, Jun Yu, and Chen Zhang. 2025. "Modeling and Forecasting Realized Volatility with Multivariate Fractional Brownian Motion." *arXiv preprint arXiv:2504.15985*.

Brandl, Florian. 2025. "The Social Learning Barrier." *arXiv preprint arXiv:2504.12136*.

Cetingoz, Adil Rengim, and Charles-Albert Lehalle. 2025. "Synthetic Data for Portfolios: A Throw of the Dice Will Never Abolish Chance." *arXiv preprint arXiv:2501.03993*.

De Oliveira, Henrique, and Rohit Lamba. 2025. "Rationalizing dynamic choices." *arXiv preprint arXiv:2504.05251*.

Fabre, Timothée, and Damien Challet. 2025. "Learning the Spoofability of Limit Order Books With Interpretable Probabilistic Neural Networks." *arXiv preprint arXiv:2504.15908*.

Gyotoku, Yoshihiro, Ioane Muni Toke, and Nakahiro Yoshida. 2025. "Deep learning of point processes

for modeling high-frequency data." *arXiv preprint arXiv:2504.15944*.

Hanauer, Matthias Xaver, Amar Soebhag, Marc Stam, and Tobias Hoogteijling. 2025. "Do Machine Learning Models Need to Be Sector Experts?" *SSRN Working Paper*.

Ji, Wenlong, Weizhe Yuan, Emily Getzen, Kyunghyun Cho, Michael I Jordan, Song Mei, Jason E Weston, Weijie J Su, Jing Xu, and Linjun Zhang. 2025. "An Overview of Large Language Models for Statisticians." *arXiv preprint arXiv:2502.17814*.

Kempf, Elisabeth, Mancy Luo, and Margarita Tsoutsoura. 2025. "The Political Economy of Firm Networks: CEO Ideology and Global Trade." *SSRN Working Paper*.

Li, Yifan, Ingmar Nolte, Sandra Nolte, and Shifan Yu. 2025. "Realized Candlestick Wicks." *Journal of Econometrics*, forthcoming.

Lopez-Lira, Alejandro, Yuehua Tang, and Mingyin Zhu. 2025. "The Memorization Problem: Can We Trust LLMs' Economic Forecasts?" *SSRN Working Paper*.

Mo, Hongwei. 2025. "Cross-Industry Technological Information Spillovers in Analyst Research." *SSRN Working Paper*.

Olea, José Luis Montiel, Mikkel Plagborg-Møller, Eric Qian, and Christian K Wolf. 2025. "Local Projections or VARs? A Primer for Macroeconomists."

Taddei, Nicola, Riccardo Maggioni, Jaap Eising, Giulia De Pasquale, and Florian Dorfler. 2025. "Distributed Time-Varying Gaussian Regression via Kalman Filtering." *arXiv preprint arXiv:2504.14900*.

Yang, Aoxiang. 2025. "Volatility-Managed Volatility Selling." *SSRN Working Paper*.

Zhang, Yichi, Mihai Cucuringu, Alexander Y Shestopaloff, and Stefan Zohren. 2025. "ClusterLOB: Enhancing Trading Strategies by Clustering Orders in Limit Order Books." *arXiv preprint arXiv:2504.20349*.