VOL. 4, NO. 5

OMI Research Newsletter – February 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.*

Announcement

The Machine Learning in Quantitative Finance Conference 2025 will be held on 5–6 June 2025 at the Mathematical Institute, University of Oxford.

Highlights

Key contributions in this month's edition include a new importance sampling scheme for cross validation in structural Bayesian hierarchical models (Han and Gelman, 2025), a novel class of multivariate heavy-tailed distributions that capture key stylised facts in economic variables and enhance statistical inference (Hansen and Tong, 2025), an equilibrium model demonstrating how market makers break anonymity to signal and collude, and affect order flow and spreads (Cartea, Chang and Graumans, 2025), an econometric framework that evaluates the capacities and limitations of large language models in empirical research (Ludwig, Mullainathan and Rambachan, 2025), the economic trade-offs between sparse models (variable selection) and dense models (shrinkage) in constructing optimal parametric portfolios from a large set of firm characteristics (Bianchi and Venturi, 2024), a statistical test for market inefficiencies caused by transitory pricing errors (Li and Yang, 2025), an adaptive-robust portfolio optimisation strategy that continuously learns unknown drift parameters of asset prices while maintaining robustness against model misspecification (Bhudisaksang, Cartea and Sánchez-Betancourt, 2025), and a continuous-time model for optimal market making with both informed and uninformed traders in the presence of short-term fads—temporary deviations from fundamental value (Barucci, Mathieu and Sánchez-Betancourt, 2025).

Machine Learning & Statistics

No Equations Needed: Learning System Dynamics Without Relying on Closed-Form ODEs

K. KACPRZYK, M. VAN DER SCHAAR
UNIVERSITY OF CAMBRIDGE

Data-driven modeling of dynamical systems is a crucial area of machine learning. In many scenarios, a thorough understanding of the model's behavior becomes essential for practical applications. For instance, understanding the behavior of a pharmacokinetic model, constructed as part of drug development, may allow us to both verify its biological plausibility (e.g., the drug concentration curve is non-negative and decays to zero) and to design dosing guidelines. Discovery of closed-form ordinary differential equations (ODEs) can be employed to obtain such insights by finding a compact mathematical equation and then analyzing it (a two-step approach). However, its widespread use is currently hindered because the analysis process may be time-consuming, requiring substantial mathematical expertise, or even impossible if the equation is too complex. Moreover, if the found equation's behavior does not satisfy the requirements, editing it or influencing the discovery algorithms to rectify it is challenging as the link between the symbolic form of an ODE and its behavior can be elusive. This paper proposes a conceptual shift to modeling low-dimensional dynamical systems by departing from the traditional two-step modeling process. Instead of first discovering a closed-form equation and then analyzing it, our approach, direct semantic modeling, predicts the semantic representation of the dynamical system (i.e., description of its behavior) directly from data, bypassing the need for complex post-hoc analysis. This direct approach also allows the incorporation of intuitive inductive biases into the optimization algorithm and editing the model's behavior directly, ensuring that the

model meets the desired specifications. Our approach not only simplifies the modeling pipeline but also enhances the transparency and flexibility of the resulting models compared to traditional closed-form ODEs.

Adaptive Sequential Monte Carlo for Automated Cross Validation in Structural Bayesian Hierarchical Models

G. Han, A. Gelman
Columbia University

Importance sampling (IS) is widely used for approximate Bayesian cross validation (CV) due to its efficiency, requiring only the re-weighting of a single set of posterior draws. With structural Bayesian hierarchical models, vanilla IS can produce unreliable results, as out-of-sample replication may involve non-standard case-deletion schemes which significantly alter the posterior geometry. This inevitably necessitates computationally expensive re-runs of Markov chain Monte Carlo (MCMC), making structural CV impracticable. To address this challenge, we consider sampling from a sequence of posteriors leading to the case-deleted posterior(s) via adaptive sequential Monte Carlo (SMC). We design the sampler to (a) support a broad range of structural CV schemes, (b) enhance efficiency by adaptively selecting Markov kernels, intervening in parallelizable MCMC re-runs only when necessary, and (c) streamline the workflow by automating the design of intermediate bridging distributions. Its practical utility is demonstrated through three real-world applications involving three types of predictive model assessments: leave-group-out CV, group K-fold CV, and sequential one-step-ahead validation.

FlashRNN: Optimizing Traditional RNNs on Modern Hardware

K. Pöppel, M. Beck, S. Hochreiter
Johannes Kepler University, NXAI Lab and NXAI GmbH

While Transformers and other sequence-parallelizable neural network architectures seem like the current state of the art in sequence modeling, they specifically lack state-tracking capabilities. These are important for time-series tasks and logical reasoning. Traditional RNNs like LSTMs and GRUs, as well as modern variants like sLSTM do have these capabilities at the cost of strictly sequential processing. While this is often seen as a strong limitation, we show how fast these networks can get with our hardware-optimization FlashRNN in Triton and CUDA, optimizing kernels to the register level on modern GPUs. We extend traditional RNNs with a parallelization variant that processes multiple RNNs of smaller hidden state in parallel, similar to the head-wise processing in Transformers. To enable flexibility on different GPU variants, we introduce a new optimization framework for hardware-internal cache sizes, memory and compute handling. It models the hardware in a setting using polyhedral-like constraints, including the notion of divisibility. This speeds up the solution process in our ConstrINT library for general integer constraint satisfaction problems (integer CSPs). We show that our kernels can achieve 50x speed-ups over a vanilla PyTorch implementation and allow 40x larger hidden sizes compared to our Triton implementation.

SPADE: Split Peak Attention DEcomposition

M. Wolff, K. G. Olivares, B. Oreshkin, S. Ruan, S. Yang, A. Katoch, S. Ramasubramanian, Y. Zhang, M. W. Mahoney, D. Efimov, V. Quenneville-Bélair
Amazon

Demand forecasting faces challenges induced by Peak Events (PEs) corresponding to special periods such as promotions and holidays. Peak events create significant spikes in demand followed by demand ramp down periods. Neural networks like MQCNN and MQT overreact to demand peaks by carrying over the elevated PE demand into subsequent Post-Peak-Event (PPE) periods, resulting in significantly over-biased forecasts. To tackle this challenge, we introduce a neural forecasting model called Split Peak Attention DEcomposition, SPADE. This model reduces the impact of PEs on subsequent forecasts by modeling forecasting as consisting of two separate tasks: one for PEs; and the other for the rest. Its architecture then uses masked convolution filters and a specialized Peak Attention module. We show SPADE's performance on a worldwide retail dataset with hundreds of millions of products. Our results reveal a reduction in PPE degradation by 4.5% and an improvement in PE accuracy by 3.9%, relative to current production models.

Tensor Product Attention Is All You Need

Y. Zhang, Y. Liu, H. Yuan, Z. Qin, Y. Yuan, Q. Gu, A. C.-C. Yao
singhua University, Shanghai Qi Zhi Institute, University of California, Los Angeles, TapTap

Scaling language models to handle longer input sequences typically necessitates large key-value (KV) caches, resulting in substantial memory overhead during inference. In this paper, we propose Tensor Product Attention (TPA), a novel attention mechanism that uses tensor decompositions to represent queries, keys, and values compactly, significantly shrinking KV cache size at inference time. By factorizing these representations into contextual low-rank components (contextual factorization) and seamlessly integrating with RoPE, TPA achieves improved model quality alongside memory efficiency. Based on TPA, we introduce the Tensor ProducT ATTenTion Transformer (T6), a new model architecture for sequence modeling. Through extensive empirical

evaluation of language modeling tasks, we demonstrate that T6 exceeds the performance of standard Transformer baselines including MHA, MQA, GQA, and MLA across various metrics, including perplexity and a range of renowned evaluation benchmarks. Notably, TPAs memory efficiency enables the processing of significantly longer sequences under fixed resource constraints, addressing a critical scalability challenge in modern language models. The code is available at this https URL.

Convolution-t Distributions

P. R. Hansen, C. Tong

University of North Carolina at Chapel Hill, Xiamen University

This paper introduces a novel class of multivariate heavy-tailed distributions, termed convolution-t distributions, which are derived from the convolutions of heterogeneous multivariate t-distributions. Unlike commonly used heavy-tailed distributions, these distributions capture nonlinear dependencies, accommodate heterogeneous marginal distributions, and reveal cluster structures, that are prevalent in economic variables. Importantly, convolution-t distributions have simple density functions that facilitate estimation and likelihood-based inference. The characteristic features of convolution-t distributions are found to be important in an empirical analysis of realized volatility measures and help identify their underlying factor structure.

Finance and Economics

Anonymity, Signaling, and Collusion in Limit Order Books

Á. Cartea, P. Chang, R. Graumans

University of Oxford

A key feature in the design of a limit order book is the anonymity of limit orders. However, we analyze data with trader identification and find that market makers break the anonymity of limit orders. Market makers use limit orders with large volumes to signal themselves to other market makers to avoid trading with each other and to snipe retail limit orders. We explain the behavior of market makers with a model that considers competitive and collusive equilibria. The model shows that the behavior of market makers we observe in the data is consistent with that in a collusive equilibrium where market makers use signals to avoid sniping each other's limit orders. Signaling enables market makers to share the benign flow from retail limit orders, and to share the additional benign flow from impatient investors who otherwise would have traded with a retail investor's limit order.

Large Language Models: An Applied Econometric Framework

J. Ludwig, S. Mullainathan, A. Rambachan

University of Chicago, Massachusetts Institute of Technology, NBER

How can we use the novel capacities of large language models (LLMs) in empirical research? And how can we do so while accounting for their limitations, which are themselves only poorly understood? We develop an econometric framework to answer this question that distinguishes between two types of empirical tasks. Using LLMs for prediction problems (including hypothesis generation) is valid under one condition: no "leakage" between the LLM's training dataset and the researcher's sample. No leakage can be ensured by using open-source LLMs with documented training data and published weights. Using LLM outputs for estimation problems to automate the measurement of some economic concept (expressed either by some text or from human subjects) requires the researcher to collect at least some validation data: without such data, the errors of the LLM's automation cannot be assessed and accounted for. As long as these steps are taken, LLM outputs can be used in empirical research with the familiar econometric guarantees we desire. Using two illustrative applications to finance and political economy, we find that these requirements are stringent; when they are violated, the limitations of LLMs now result in unreliable empirical estimates. Our results suggest the excitement around the empirical uses of LLMs is warranted – they allow researchers to effectively use even small amounts of language data for both prediction and estimation – but only with these safeguards in place.

Rethinking Sparsity: Parametric Portfolios and Firm Characteristics

D. Bianchi and P. Venturi

Queen Mary University of London

We investigate the economic trade-off between variable selection and shrinkage in designing optimal parametric portfolios based on a large set of firm characteristics. Using a flexible Bayesian inference approach, we show that a sparsity-inducing prior reduces model uncertainty but leads to high-turnover, under-diversified portfolios with concentrated exposures to a few characteristics. In contrast, a prior favoring ridge-type shrinkage fosters better diversification across more characteristics, yielding portfolios more aligned with mean-variance efficiency, particularly when accounting for realistic transaction costs. Our findings caution against over-reliance on sparsity-inducing methods when predicting the cross-section of stock returns using firm characteristics.

Realized Weight Regression for Dynamic Real-

ized Global Minimum Variance Portfolio Allocation

E. Kazak and Y. Li and I. Nolte and S. Nolte
University of Birmingham, University of Manchester, Lancaster University

This paper introduces the realized weight regression (RWR), a novel and easy-to-implement method for dynamic realized global minimum variance portfolio (GVMP) allocation. The RWR approach minimizes the unconditional expected realized portfolio variance within a class of linear weight models, which can be interpreted as a linear regression of the transformed realized GMVP weight vectors on a set of predictive weights. We prove that, under mild technical conditions, the RWR parameters can be consistently estimated by least squares for an arbitrarily large number of assets in finite samples, and as such RWR do not suffer from the curse of dimensionality. RWR can be extended to incorporate controls for investment constraints to balance economic performance with transaction costs. Our empirical results demonstrate that, under different number of assets, estimation window length, and investment constraints, RWR consistently outperforms competing estimators in terms of out-of-sample portfolio variance measures with realistic portfolio weights.

Multi-Horizon Test for Market Efficiency

Z. M. Li, X. Yang
The Chinese University of Hong Kong, Rutgers University

In efficient markets, asset returns show no predictability over short periods, such as daily or intraday intervals, where expected returns are nearly zero. However, the presence of pricing errors–transitory components of asset prices reflecting various market frictions or biases–induces return reversals. These reversals provide evidence against market efficiency, particularly in terms of liquidity provision. We propose to test market efficiency based on the joint inference of the covariances of observed returns over multiple horizons. We show that analyzing a small set of horizons is sufficient to detect inefficiencies caused by a wide range of transitory pricing errors, both theoretically and practically. Extensive simulations highlight the superiority of our multi-horizon approach over traditional tests, especially when pricing errors are weak and exhibit complex serial dependencies. Moreover, our test statistic serves as a natural liquidity measure that can effectively identify financial crisis with significant liquidity drains.

Bursting Bubbles in a Macroeconomic Model

T. Hirano, K. Kishi, A. A. Toda
Royal Holloway, University of London, London School of Economic, Kansai University, Emory University

This paper identifies the conditions and mechanisms that give rise to stochastic bubbles that are expected to collapse. To illustrate the essence of the emergence of stochastic bubbles, we first present a toy model, and then we present a full-fledged macro-finance model of intangible capital and show that stochastic stock bubbles attached to intangible capital emerge in the process of spillover of technological innovation. We show that the dynamics with stochastic bubbles, which is characterized by unbalanced growth, is a temporary deviation from a balanced growth path in which asset prices equal the fundamentals.

How Much of Cross-Stock Momentum Reflects Underreaction?

J. Li, J. Yan
University of Utah, Renmin University of China

Many studies find that returns of related stocks cross-predict each other ("cross-stock momentum"). These patterns have been interpreted as underreaction to cash flow information, which implies that the returns should persist. However, we document that they revert over the long run. To further understand cross-stock momentum, we show that a significant component of these returns reflect common momentum in factor portfolios, which fully accounts for the reversal. After removing these effects, the resulting return predictability appears more consistent with underreaction: it represents uni-directional lead-lags, and the returns mainly manifest in small and hard-to-arbitrage stocks. On a dollar-weighted basis, only a small portion of cross-stock momentum appears consistent with underreaction.

High-Frequency Estimation of Itô Semimartingale Baseline for Hawkes Processes

Y. Potiron, O. Scaillet, V. Volkov, S. Yu
Keio University, University of Geneva, Swiss Finance Institute, University of Tasmania, Northwestern University

We consider Hawkes self-exciting processes with a baseline driven by an Itô semimartingale with possible jumps. Under in-fill asymptotics, we characterize feasible statistics induced by central limit theory for empirical average and variance of local Poisson estimates. As a byproduct, we develop a test for the absence of a Hawkes component and a test for baseline constancy. Simulation studies corroborate the asymptotic theory. An empirical application on high-frequency data of the E-mini SP500 future contracts shows that the absence of a Hawkes component and baseline constancy is always rejected.

Mathematical Finance

Adaptive-Robust Portfolio Optimisation

T. Bhudisaksang, Á. Cartea, L. Sánchez-Betancourt

University of Oxford

An agent solves a multidimensional investment problem that is robust to parameter misspecification and where the optimal strategy continuously adapts to new information. The agent invests in a risk-free asset and in risky stocks whose prices follow a geometric diffusion processes. The agent does not know the drift parameters of the stock price dynamics, so she considers a set of alternative measures to make the investment problem robust to model misspecification and employs a continuous-time estimator to learn the value of the drift parameters as new information arrives during the investment horizon. The agent's value function is characterised as the solution to a nonlinear PDE. We show that the value function has a stochastic representation and use it to analyse the optimal adaptive-robust strategy and to compare it with various benchmarks.

Market Making with Fads, Informed, and Uninformed Traders

E. Barucci, A. Mathieu, L. Sánchez-Betancourt

Politecnico di Milano, University of Oxford

We characterise the solutions to a continuous-time optimal liquidity provision problem in a market populated by informed and uninformed traders. In our model, the asset price exhibits fads-these are short-term deviations from the fundamental value of the asset. Conditional on the value of the fad, we model how informed traders and uninformed traders arrive in the market. The market maker knows of the two groups of traders but only observes the anonymous order arrivals. We study both, the complete information and the partial information versions of the control problem faced by the market maker. In such frameworks, we characterise the value of information, and we find the price of liquidity as a function of the proportion of informed traders in the market. Lastly, for the partial information setup, we explore how to go beyond the Kalman-Bucy filter to extract information about the fad from the market arrivals.

A Groundwater Market Model

I. Cialenco, M. Ludkovski

Illinois Institute of Technology, University of California, Santa Barbara

We introduce the problem of groundwater trading, capturing the emergent groundwater market setups among stakeholders in a given groundwater basin. The agents optimize their production, taking into account their available water rights, the requisite water consumption, and the opportunity to trade water among themselves. We study the resulting Nash equilibrium, providing a full characterization of the 1-period setting and initial results about the features of the multi-period game driven by the ability of agents to bank their water rights in order to smooth out the intertemporal shocks.

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, Qi Jin, and Harrison Waldon.

References

Barucci, Emilio, Adrien Mathieu, and Leandro Sánchez-Betancourt. 2025. "Market Making with Fads, Informed, and Uninformed Traders." *SSRN Working Paper*.

Bhudisaksang, Theerawat, Álvaro Cartea, and Leandro Sánchez-Betancourt. 2025. "Adaptive-Robust Portfolio Optimisation." *SSRN Working Paper*.

Bianchi, Daniele, and Pedro Venturi. 2024. "Rethinking Sparsity: Parametric Portfolios and Firm Characteristics." *SSRN Working Paper*.

Cartea, Álvaro, Patrick Chang, and Rob Graumans. 2025. "Anonymity, Signaling, and Collusion in Limit Order Books." *SSRN Working Paper*.

Cialenco, Igor, and Michael Ludkovski. 2025. "A Groundwater Market Model." *arXiv preprint arXiv:2501.14071*.

Han, Geonhee, and Andrew Gelman. 2025. "Adaptive Sequential Monte Carlo for Automated Cross Validation in Structural Bayesian Hierarchical Models." *arXiv preprint arXiv:2501.07685*.

Hansen, Peter Reinhard, and Chen Tong. 2025. "Convolution-t Distributions." *SSRN Working Paper*.

Hirano, Tomohiro, Keiichi Kishi, and Alexis Akira Toda. 2025. "Bursting Bubbles in a Macroeconomic Model." *arXiv preprint arXiv:2501.08215*.

Kacprzyk, Krzysztof, and Mihaela van der Schaar. 2025. "No Equations Needed: Learning System Dynamics Without Relying on Closed-Form ODEs." *arXiv preprint arXiv:2501.18563*.

Kazak, Ekaterina, Yifan Li, Ingmar Nolte, and Sandra Nolte. 2025. "Realized Weight Regression for Dynamic Realized Global Minimum Variance Portfolio Allocation." *SSRN Working Paper*.

Li, Z Merrick, and Xiye Yang. 2025. "Multi-Horizon Test for Market Efficiency." *Working Paper*.

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan. 2025. "Large Language Models: An Applied Econometric Framework." *NBER Working Paper*.

Pöppel, Korbinian, Maximilian Beck, and Sepp Hochreiter. 2025. "FlashRNN: Optimizing Traditional RNNs on Modern Hardware." *arXiv preprint arXiv:2412.07752*.

Potiron, Yoann, Olivier Scaillet, Vladimir Volkov, and Seunghyeon Yu. 2025. "High-Frequency Estimation of Itô Semimartingale Baseline for Hawkes Processes." *SSRN Working Paper*.

Wolff, Malcolm, Kin G Olivares, Boris Oreshkin, Sunny Ruan, Sitan Yang, Abhinav Katoch, Shankar Ramasubramanian, Youxin Zhang, Michael W Mahoney, Dmitry Efimov, et al. 2025. "SPADE: Split Peak Attention DEcomposition." *arXiv preprint arXiv:2411.05852*.

Zhang, Yifan, Yifeng Liu, Huizhuo Yuan, Zhen Qin, Yang Yuan, Quanquan Gu, and Andrew Chi-Chih Yao. 2025. "Tensor Product Attention Is All You Need." *arXiv preprint arXiv:2501.06425*.