OMI Research Newsletter – January 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 Oxford–Harvard Conference on Decentralised Finance will be held on 20–21 January 2025 at the Oxford-Man Institute. Please find the schedule and register here.

Highlights

Key contributions in this month's edition include a bias-corrected framework for statistical inference with AI- and ML-generated variables (Battaglia et al., 2024), a nonlinear Bayesian filter that addresses linearisation errors and achieves local convergence to the optimal Gaussian approximation (Cao et al., 2024), an investigation into the economic significance and interpretability of the inner beliefs of large language models (Chen, Didisheim and Somoza, 2024), insights into the role of noise in economic forecasting and its impact on predictive accuracy (Liao et al., 2024), the identification of the "end-of-day reversal" pattern in individual stock returns and its implications for asset pricing (Baltussen, Da and Soebhag, 2024), an analysis of strategic interactions between brokers and informed traders in broker-mediated markets (Aqsha, Drissi and Sánchez-Betancourt, 2024), and a study exploring the market impact of passive orders and its influence on liquidity and price dynamics (Chahdi, Rosenbaum and Szymanski, 2024).

Machine Learning & Statistics

Inference for Regression with Variables Generated by AI or Machine Learning

L. BATTAGLIA, T. CHRISTENSEN, S. HANSEN, S. SACHER

UNIVERSITY OF OXFORD, YALE UNIVERSITY, UNIVERSITY COLLEGE LONDON, META

It has become common practice for researchers to use AI-powered information retrieval algorithms or other machine learning methods to estimate variables of economic interest, then use these estimates as covariates in a regression model. We show both theoretically and empirically that naively treating AI- and ML-generated variables as "data" leads to biased estimates and invalid inference. We propose two methods to correct bias and perform valid inference: (i) an explicit bias correction with bias-corrected confidence intervals, and (ii) joint maximum likelihood estimation of the regression model and the variables of interest. Through several applications, we demonstrate that the common approach generates substantial bias, while both corrections perform well.

Nonlinear Bayesian Filtering with Natural Gradient Gaussian Approximation

W. CAO, T. ZHANG, Z. SUN, C. LIU, S. S.-T. YAU, S. E. LI

TSINGHUA UNIVERSITY, PEKING UNIVERSITY

Practical Bayes filters often assume the state distribution of each time step to be Gaussian for computational tractability, resulting in the so-called Gaussian filters. When facing nonlinear systems, Gaussian filters such as extended Kalman filter (EKF) or unscented Kalman filter (UKF) typically rely on certain linearization techniques, which can introduce large estimation errors. To address this issue, this paper reconstructs the prediction and update steps of Gaussian filtering as solutions to two distinct optimization problems, whose optimal conditions are found to have analytical forms from Stein's lemma.

It is observed that the stationary point for the prediction step requires calculating the first two moments of the prior distribution, which is equivalent to that step in existing moment-matching filters. In the update step, instead of linearizing the model to approximate the stationary points, we propose an iterative approach to directly minimize the update step's objective to avoid linearization errors. For the purpose of performing the steepest descent on the Gaussian manifold, we derive its natural gradient that leverages Fisher information matrix to adjust the gradient direction, accounting for the curvature of the parameter space. Combining this update step with moment matching in the prediction step, we introduce a new iterative filter for nonlinear systems called *N*atural Gr*a*dient Gaussi*a*n App*r*oximation filter, or NANO filter for short. We prove that NANO filter locally converges to the optimal Gaussian approximation at each time step. Furthermore, the estimation error is proven exponentially bounded for nearly linear measurement equation and low noise levels through constructing a supermartingale-like property across consecutive time steps.

Streamlining Prediction in Bayesian Deep Learning

R. Li, M. Klasson, A. Solin, M. Trapp
Aalto University

The rising interest in Bayesian deep learning (BDL) has led to a plethora of methods for estimating the posterior distribution. However, efficient computation of inferences, such as predictions, has been largely overlooked with Monte Carlo integration remaining the standard. In this work we examine streamlining prediction in BDL through a single forward pass without sampling. For this we use local linearisation on activation functions and local Gaussian approximations at linear layers. Thus allowing us to analytically compute an approximation to the posterior predictive distribution. We showcase our approach for both MLP and transformers, such as ViT and GPT-2, and assess its performance on regression and classification tasks.

Relaxed Recursive Transformers: Effective Parameter Sharing with Layer-wise LoRA

S. Bae, A. Fisch, H. Harutyunyan, Z. Ji, S. Kim, T. Schuster
KAIST AI, Google

Large language models (LLMs) are expensive to deploy. Parameter sharing offers a possible path towards reducing their size and cost, but its effectiveness in modern LLMs remains fairly limited. In this work, we revisit "layer tying" as form of parameter sharing in Transformers, and introduce novel methods for converting existing LLMs into smaller "Recursive Transformers" that share parameters across layers, with minimal loss of performance. Here, our Recursive Transformers are efficiently initialized from standard pretrained Transformers, but only use a single block of unique layers that is then repeated multiple times in a loop. We further improve performance by introducing Relaxed Recursive Transformers that add flexibility to the layer tying constraint via depth-wise low-rank adaptation (LoRA) modules, yet still preserve the compactness of the overall model. We show that our recursive models (e.g., recursive Gemma 1B) outperform both similar-sized vanilla pretrained models (such as TinyLlama 1.1B and Pythia 1B) and knowledge distillation baselines – and can even recover most of the performance of the original "full-size" model (e.g., Gemma 2B with no shared parameters). Finally, we propose Continuous Depth-wise Batching, a promising new inference paradigm enabled by the Recursive Transformer when paired with early exiting. In a theoretical analysis, we show that this has the potential to lead to significant (2-3x) gains in inference throughput.

Byte Latent Transformer: Patches Scale Better Than Tokens

A. Pagnoni, R. Pasunuru, P. Rodriguez, J. Nguyen, B. Muller, M. Li, C. Zhou, L. Yu, J. Weston, L. Zettlemoyer, G. Ghosh, M. Lewis, A. Holtzman, S. Iyer
Meta, University of Washington, University of Chicago

We introduce the Byte Latent Transformer (BLT), a new byte-level LLM architecture that, for the first time, matches tokenization-based LLM performance at scale with significant improvements in inference efficiency and robustness. BLT encodes bytes into dynamically sized patches, which serve as the primary units of computation. Patches are segmented based on the entropy of the next byte, allocating more compute and model capacity where increased data complexity demands it. We present the first FLOP controlled scaling study of byte-level models up to 8B parameters and 4T training bytes. Our results demonstrate the feasibility of scaling models trained on raw bytes without a fixed vocabulary. Both training and inference efficiency improve due to dynamically selecting long patches when data is predictable, along with qualitative improvements on reasoning and long tail generalization. Overall, for fixed inference costs, BLT shows significantly better scaling than tokenization-based models, by simultaneously growing both patch and model size.

Out of the (Black)Box: AI as Conditional Probability

H. Chen, A. Didisheim, L. Somoza
Massachusetts Institute of Technology, University of Melbourne, ESSEC Business School.

We explore the economic significance and interpretability of the distribution of conditional probabilities behind LLM's text generation. Using a dataset of news and returns, we find that conditional probabilities are interpretable and correlate with model accuracy. Conversely, measures of declared confidence used in the literature are opaque, structurally biased, unstable, and more model-dependent, indicating that LLMs cannot assess their own confidence. Using conditional probabilities, we analyze LLM biases and provide insights into the internal mechanisms driving model decisions. Our results indicate that conditional probabilities provide a reliable and transparent reflection of LLM beliefs, particularly for economic applications.

Finance and Economics

Does Finance Benefit Society? A Language Embedding Approach

M. Jha, H. Liu, A. Manela

Georgia State University, Washington University in St. Louis

We measure popular sentiment toward finance by applying a large language model to millions of books published in eight countries over hundreds of years. We extensively validate this measure both internally and externally. We document persistent differences in finance sentiment across countries despite ample time-series variation. Books written in the languages of more capitalist countries discuss finance in a more positive context. Finance sentiment is correlated with survey-based measures of financial market participation and income inequality. Finance sentiment declines one year before rather than after financial crises. Positive shocks to finance sentiment are followed by higher output and credit growth.

Does Noise Hurt Economic Forecasts?

Y. Liao, X. Ma, A. Neuhierl, Z. Shi

National University of Singapore, Washington University in St. Louis, Chinese University of Hong Kong

This paper explores whether variable selection enhances economic forecasting. While economists often remove noise from predictors, we show that economic forecast models are not sparse if the outcome is driven by latent factors. We also prove a compelling result that including noise in predictions yields greater benefits than excluding it. Empirically, we apply this method to four common forecasting applications including forecasting the U.S. inflation rate and obtain results that surpass many commonly used models that rely on dimension reduction or variable selection.

End-of-Day Reversal

G. Baltussen, Z. Da, A. Soebhag

Erasmus University Rotterdam, University of Notre Dame, Northern Trust Asset Management, Robeco Quantitative Investing

Individual stocks experience sharp intraday return reversals in the cross-section during the last 30 minutes of the trading day. This "end-of-day reversal" pattern is economically and statistically highly significant, is distinct from market intraday momentum, and primarily comes from positive price pressure on intraday losers. The effect cannot be explained by liquidity or gamma hedging effects. Instead, two novel channels related to the attention-induced retail purchases and risk management by short-sellers at the end of the day are driving the effect.

Momentum and Factor Momentum: A Re-Examination

C. Gao, S. Z. Li, P. Yuan, G. Zhou

Soochow University, Rutgers University, Hong Kong Baptist University, Washington University in St. Louis

We show that the momentum factor remains a unique and irreplaceable factor, in contrast to the redundancy finding of Ehsani and Linnainmaa (2022), which suffers from an omitted-variable problem. By adding a betting-against-systematic factor to their framework, we find that the momentum factor exhibits significant alpha. Further, we demonstrate that even an improved factor model, such as IPCA, cannot explain the momentum unless momentum characteristics are utilized. Moreover, in an attribution analysis, we show that firm-specific components, not non-momentum factors, are the primary drivers of momentum returns.

Design Choices, Machine Learning, and the Cross-Section of Stock Returns

M. Chen, M. X. Hanauer, T. Kalsbach

Technische Universität München, Robeco

We fit over one thousand machine learning models for predicting stock returns, systematically varying design choices across algorithm, target variable, feature selection, and training methodology. Our findings demonstrate that the non-standard error in portfolio returns arising from these design choices exceeds the standard error by 59%. Furthermore, we observe a substantial variation in model performance, with monthly mean top-minus-bottom returns ranging from 0.13% to 1.98%. These findings underscore the critical impact of design choices on machine learning predictions, and we offer recommendations for model design. Finally, we identify the conditions under which non-linear models outperform linear models.

Neural Tangent Kernel in Implied Volatility Forecasting: A Nonlinear Functional Autoregression Approach

Y. Chen, M. Grith, H. Lai
National University of Singapore, Erasmus University Rotterdam, National University of Singapore

Forecasting implied volatility across different levels of moneyness and maturity is crucial yet challenging due to the high dimensionality of the Implied Volatility Surface (IVS) and the nonlinearity that characterizes its temporal dependence. We adopt a Nonlinear Functional Autoregressive (NFAR) framework to a sequence of IVS and employ neural networks that admit a Neural Tangent Kernel (NTK) parametrization to capture nonlinear interactions between surfaces. We illustrate the theoretical and numerical advantages of the proposed functional NTK (fNTK) estimator and establish a link to functional kernel regression. Our empirical analysis includes over 6 million European calls and put options from the SP 500 Index, covering January 2009 to December 2021. The results confirm the superior forecasting accuracy of the fNTK across different time horizons. When applied to short delta-neutral straddle trading, the fNTK achieves a Sharpe ratio ranging from 1.30 to 1.83 on a weekly to monthly basis, translating to 90% to 675% relative improvement in portfolio returns compared to forecasts based on functional Random Walk model.

Strategic Attribute Learning

J.-M. Benkert, L. Matyskova, E. Starkov
University of Bern, University of Alicante, University of Copenhagen

A researcher allocates a budget of informative tests across multiple unknown attributes to influence a decision-maker. We derive the researcher's equilibrium learning strategy by solving an auxiliary single-player problem. The attribute weights in this problem depend on how much the researcher and the decision-maker disagree. If the researcher expects an excessive response to new information, she forgoes learning altogether. In an organizational context, we show that a manager favors more diverse analysts as the hierarchical distance grows. In another application, we show how an appropriately opposed advisor can constrain a discriminatory politician, and identify the welfare-inequality Pareto frontier of researchers.

Mathematical Finance

Strategic Learning and Trading in Broker-Mediated Markets

A. Aqsha, F. Drissi, L. Sánchez-Betancourt
University of Oxford

We study strategic interactions in a broker-mediated market. A broker provides liquidity to an informed trader and to noise traders while managing inventory in the lit market. The broker and the informed trader maximise their trading performance while filtering each other's private information; the trader estimates the broker's trading activity in the lit market while the broker estimates the informed trader's private signal. Brokers hold a strategic advantage over traders who rely solely on prices to filter information. We find that information leakage in the client's trading flow yields an economic value to the broker that is comparable to transaction costs; she speculates profitably and mitigates risk effectively, which, in turn, adversely impacts the informed trader's performance. In contrast, low signal-to-noise sources, such as prices, result in the broker's trading performance being indistinguishable from that of a naive strategy that internalises noise flow, externalises informed flow, and offloads inventory at a constant rate.

Broker-Trader Partial Information Nash Equilibria

X. Wu, S. Jaimungal
University of Toronto

We study partial information Nash equilibrium between a broker and an informed trader. In this model, the informed trader, who possesses knowledge of a trading signal, trades multiple assets with the broker in a dealer market. Simultaneously, the broker trades these assets in a lit exchange where their actions impact the asset prices. The broker, however, only observes aggregate prices and cannot distinguish between underlying trends and volatility. Both the broker and the informed trader aim to maximize their penalized expected wealth. Using convex analysis, we characterize the Nash equilibrium and demonstrate its existence and uniqueness. Furthermore, we establish that this equilibrium corresponds to the solution of a nonstandard system of forward-backward stochastic differential equations.

A Theory of Passive Market Impact

Y. O. Chahdi, M. Rosenbaum, G. Szymanski
CentraleSupélec, École polytechnique, University of Luxembourg

While the market impact of aggressive orders has been extensively studied, the impact of passive orders, those executed through limit orders, remains less understood. The goal of this paper is to investigate passive market impact by developing a microstructure model connecting liquidity dynamics and price moves. A key innovation of our approach is to replace the traditional assumption of constant information content for each trade by a function that depends on the available volume in the limit order

book. Within this framework, we explore scaling limits and analyze the market impact of passive metaorders. Additionally, we derive useful approximations for the shape of market impact curves, leading to closed-form formulas that can be easily applied in practice.

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

Aqsha, Alif, Fayçal Drissi, and Leandro Sánchez-Betancourt. 2024. "Strategic Learning and Trading in Broker-Mediated Markets." *arXiv preprint arXiv:2412.20847*.

Bae, Sangmin, Adam Fisch, Hrayr Harutyunyan, Ziwei Ji, Seungyeon Kim, and Tal Schuster. 2024. "Relaxed Recursive Transformers: Effective Parameter Sharing with Layer-wise LoRA." *arXiv preprint arXiv:2410.20672*.

Baltussen, Guido, Zhi Da, and Amar Soebhag. 2024. "End-of-Day Reversal." *Available at SSRN 5039009*.

Battaglia, Laura, Timothy Christensen, Stephen Hansen, and Szymon Sacher. 2024. "Inference for Regression with Variables Generated by AI or Machine Learning." *arXiv preprint arXiv:2402.15585*.

Benkert, Jean-Michel, Ludmila Matyskova, and Egor Starkov. 2024. "Strategic Attribute Learning." *arXiv preprint arXiv:2412.10024*.

Cao, Wenhan, Tianyi Zhang, Zeju Sun, Chang Liu, Stephen S-T Yau, and Shengbo Eben Li. 2024. "Nonlinear Bayesian Filtering with Natural Gradient Gaussian Approximation." *arXiv preprint arXiv:2410.15832*.

Chahdi, Youssef Ouazzani, Mathieu Rosenbaum, and Grégoire Szymanski. 2024. "A Theory of Passive Market Impact." *arXiv preprint arXiv:2412.07461*.

Chen, Hui, Antoine Didisheim, and Luciano Somoza. 2024. "Out of the (Black) Box: AI as Conditional Probability." *Available at SSRN 5012852*.

Chen, Minghui, Matthias X Hanauer, and Tobias Kalsbach. 2024. "Design Choices, Machine Learning, and the Cross-Section of Stock Returns." *Available at SSRN 5031755*.

Chen, Ying, Maria Grith, and Hannah Lai. 2024. "Neural Tangent Kernel in Implied Volatility Forecasting: A Nonlinear Functional Autoregression Approach." *Available at SSRN 4547560*.

Gao, Cheng, Sophia Zhengzi Li, Peixuan Yuan, and Guofu Zhou. 2024. "Momentum and Factor Momentum: A Re-Examination." *Available at SSRN 5057525*.

Jha, Manish, Hongyi Liu, and Asaf Manela. 2024. "Does Finance Benefit Society? A Language Embedding Approach." *Review of Financial Studies*, forthcoming.

Liao, Yuan, Xinjie Ma, Andreas Neuhierl, and Zhentao Shi. 2024. "Does Noise Hurt Economic Forecasts?" *Available at SSRN 4659309*.

Li, Rui, Marcus Klasson, Arno Solin, and Martin Trapp. 2024. "Streamlining Prediction in Bayesian Deep Learning." *arXiv preprint arXiv:2411.18425*.

Pagnoni, Artidoro, Ram Pasunuru, Pedro Rodriguez, John Nguyen, Benjamin Muller, Margaret Li, Chunting Zhou, Lili Yu, Jason Weston, Luke Zettlemoyer, et al. 2024. "Byte Latent Transformer: Patches Scale Better Than Tokens." *arXiv preprint arXiv:2412.09871*.

Wu, Xuchen, and Sebastian Jaimungal. 2024. "Broker-Trader Partial Information Nash Equilibria." *arXiv preprint arXiv:2412.17712*.