VOL. 5, NO. 8

OMI Research Newsletter – May 2026

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 OMI Machine Learning and Finance Conference will be held at the Mathematical Institute on 11–12 June 2026. For the schedule and registration details, follow this link.

The OMI–SBS Conference on FinTech and Blockchain Economics will be held at Saïd Business School and OMI on 25–26 June 2026. For the schedule and registration details, follow this link.

Highlights

Key contributions in this month's edition include the following. Yiu et al. (2026) derive a robust update rule for the online infinite hidden Markov model (iHMM) when streaming data contain outliers and the model is misspecified. Ge et al. (2026) use text-derived priors to guide LASSO penalties, improving high-dimensional variable selection and financial prediction. Christensen et al. (2026) develop a high-frequency, nonparametric realised copula to measure dependence between latent volatility processes. Didisheim et al. (2026) show that stock markets substantially misprice "news shocks"—the unpredictable component of news text after removing content predictable from firm characteristics. Timmermann and Vulicevic (2026) show that return predictability follows compute scaling laws, with additional compute systematically improving machine-learning portfolio performance towards an estimable limit. Babiak, Barunik and Kurka (2026) show that cross-sectional dispersion in firm-level realised skewness strongly and negatively predicts future market returns. Ouyang and Sui (2026) show that LLM trading agents reproduce human behavioural biases, generate market bubbles, and can be stabilised through targeted prompt-based guardrails. Finally, Noble, Rosenbaum and Souilmi (2026) build an interactive limit-order-book simulator that improves queue-reactive models by incorporating realistic timing, latency races, and market-impact feedback.

Statistics, Econometrics & Machine Learning

Doubly Outlier-Robust Online Infinite Hidden Markov Model

H. YIU, L. SÁNCHEZ-BETANCOURT, Á. CARTEA, G. DURAN-MARTIN

OXFORD-MAN INSTITUTE, UNIVERSITY OF OXFORD

We derive a robust update rule for the online infinite hidden Markov model (iHMM) for when the streaming data contains outliers and the model is misspecified. Leveraging recent advances in generalised Bayesian inference, we define robustness via the posterior influence function (PIF), and provide conditions under which the online iHMM has bounded PIF. Imposing robustness inevitably induces an adaptation lag for regime switching. Our method, which is called Batched Robust iHMM (BR-iHMM), balances adaptivity and robustness with two additional tunable parameters. Across limit order book data, hourly electricity demand, and a synthetic high-dimensional linear system, BR-iHMM reduces one-step-ahead forecasting error by up to 67% relative to competing online Bayesian methods. Together with theoretical guarantees of bounded PIF, our results highlight the practicality of our approach for both forecasting and interpretable online learning.

Text as Priors

S. Ge, S. Li, O. B. Linton, W. Su

Nankai University, Peking University, University of Cambridge, University of Oxford

This paper studies how textual information can be used as prior information in high-dimensional penalized estimation. Rather than treating text as an additional set of regressors, we use textual analysis to construct two classes of priors: linkage priors, which encode the relative relevance of covariates, and direction priors, which encode prior information about coefficient signs. We incorporate these priors through weighted and asymmetric LASSO procedures. We show that, when the priors are sufficiently informative, they improve variable selection by relaxing the irrepresentable condition required for selection consistency of the standard LASSO, especially in settings with strongly correlated covariates. We illustrate the framework in three applications based on Chinese financial news: cross-firm return prediction, large precision matrix estimation for portfolio construction, and high-dimensional text regression with sentiment-based sign restrictions. Overall, the results show that text can enhance high-dimensional estimation not only as data, but also as a source of economically meaningful prior information.

The Realized Copula of Volatility

K. Christensen, W. Liu, Z. Liu, Y. Potiron

Aarhus University, University of Macau, Keio University

We study a new measure of codependency in the second moment of a continuous-time multivariate asset price process, which we name the realized copula of volatility. The statistic is based on local volatility estimates constructed from high-frequency asset returns and affords a nonparametric estimator of the empirical copula of the latent stochastic volatility. We show consistency of our estimator with in-fill asymptotic theory, either with a fixed or increasing time span. In the latter setting, we derive a functional central limit theorem for the empirical process associated with the measurement error of the time-invariant marginal copula of volatility. We also develop a goodness-of-fit test to evaluate hypotheses about the shape of the latter. In a simulation study, we demonstrate that our estimator is a good proxy of both the empirical and marginal copula of volatility, even with a moderate amount of high-frequency data recorded over a relatively short sample. The goodness-of-fit test is found to exhibit size control and excellent power. We implement our framework on high-frequency transaction data from futures contracts that track the U.S. equity and treasury bond market. A Gumbel copula is found to offer a near-perfect bind between the realized variance processes in these data.

An Introduction to Double/Debiased Machine Learning

A. Ahrens, V. Chernozhukov, C. Hansen, D. Kozbur, M. Schaffer, T. Wiemann

This paper provides an introduction to Double/Debiased Machine Learning (DML). DML is a general approach to performing inference about a target parameter in the presence of nuisance functions: objects that are needed to identify the target parameter but are not of primary interest. Nuisance functions arise naturally in many settings, such as when controlling for confounding variables or leveraging instruments. The paper describes two biases that arise from nuisance function estimation and explains how DML alleviates these biases. Consequently, DML allows the use of flexible methods, including machine learning tools, for estimating nuisance functions, reducing the dependence on auxiliary functional form assumptions and enabling the use of complex non-tabular data, such as text or images. We illustrate the application of DML through simulations and empirical examples. We conclude with a discussion of recommended practices. A companion website includes additional examples with code and references to other resources.

Testing Alpha in High-Dimensional Conditional Time-Varying Factor Models with Dependent Observations

L. Feng, H. Ma, Z. Wang

Nankai University

This paper studies alpha testing in a high-dimensional conditional time-varying factor model with temporally dependent observations. Both factor loadings and alpha processes are allowed to vary smoothly over time, and the cross-sectional dimension may be comparable to or larger than the sample size. Using a B-spline sieve method, we develop a sum-type test for dense alternatives, a max-type test for sparse alternatives, and a Cauchy combination test for adaptive inference. On the theoretical side, we derive explicit stochastic expansions for the estimated average alphas, establish asymptotic normality of the sum statistic, and develop the extreme-value limit theory for the max statistic by showing its Gumbel convergence under temporal dependence together with the validity of block-bootstrap calibration. We further prove asymptotic independence between the sum and max statistics and thereby justify the Cauchy combination test. Simulation results demonstrate that the proposed procedures achieve satisfactory size control and competitive power across a wide range of dense and sparse alternatives. An empirical application further illustrates the usefulness of the proposed methods in testing asset-pricing models with time-varying structure.

Generalized Autoregressive Multivariate Models: From Binary to Poisson

A. Bykhovskaya, N. Meddahi
Duke University, Toulouse School of Economics

This paper presents a framework for binary autoregressive time series in which each observation is a Bernoulli variable whose success probability evolves with past outcomes and probabilities, in the spirit of GARCH-type dynamics, accommodating nonlinearities, network interactions, and cross-sectional dependence in the multivariate case. Existence and uniqueness of a stationary solution is established via a coupling argument tailored to the discontinuities inherent in binary data. A key theoretical result, further supported by our empirical illustration on S&P 100 data, shows that, under a rare-events scaling, aggregates of such binary processes converge to a Poisson autoregression, providing a micro-foundation for this widely used count model. Maximum likelihood estimation is proposed and illustrated empirically.

Finance and Economics

The Inefficient Pricing of News

A. Didisheim, B. T. Kelly, M. Pourmohammadi, H. Tian
University of Melbourne, Yale University

The stock market fails to efficiently process information in news text (Chen et al., 2026). But news itself is highly predictable by prevailing stock characteristics, which complicates inferences about market efficiency. After purging news of its predictable content, the resulting "news shocks" more than double the monthly return predictive power of raw news, and they continue to significantly predict returns up to 18 months ahead. The magnitude and longevity of the news shock anomaly is larger than every anomaly in the Jensen et al. (2022) universe. The news shock anomaly derives from negative-tone and quantitative topics to which investors underreact and from high-attention and ambiguous topics to which investors overreact.

Compute, Complexity, and the Scaling Laws of Return Predictability

A. Timmermann, L. Vulicevic
University of California, San Diego

Computing power ("compute") governs how much signal investors can extract from high-dimensional data. Using firm-level cross-sectional data, we find that return predictability follows stable power-law relationships between accuracy and the compute expended on model training, explaining over 80% of the variation in portfolio performance across models. These scaling laws jointly identify the limit to return predictability and the rate at which additional compute closes the gap to that limit. Compared to complexity, compute stands out as a scarce capital resource that bottlenecks how many parameters one can effectively use. By treating compute as a primitive, scaling laws can quantify the value of new data, distinguish problems with strong versus weak returns to scale, and finally, measure the elasticity of risk-adjusted return performance with respect to compute. Under this metric, a 25% increase in compute would have raised an investor's Sharpe ratio by about 14% over the past three decades, indicating substantial returns to computational sophistication.

Which Factors Drive Downside Risk in the U.S. Economy?

C. T. Brownlees, C. Pavanello, A. B. M. Souza
Universitat Pompeu Fabra, ESADE Business School

We study which common factors drive downside risk across a large panel of U.S. macroeconomic variables. We consider a broad set of candidate predictors, comprising both observed factors constructed from macroeconomic, financial, and text data, as well as unobserved factors associated with the panel. The relevance of the factors is assessed by how much they improve out-of-sample downside risk prediction accuracy. Factors are mapped into forecasts via quantile regression and location-scale regression. Results point to a single factor associated with macroeconomic volatility, most closely proxied by the macroeconomic uncertainty index (Jurado et al., 2015).

Skewness Dispersion and Stock Market Returns

M. Babiak, J. Barunik, J. Kurka
Lancaster University, Charles University

Cross-sectional dispersion in firm-level realized skewness is significantly and negatively related to future stock market returns. The predictive power of skewness dispersion is robust to in-sample and out-of-sample estimation and is incremental over a broad set of existing predictors, with only a few alternatives retaining independent explanatory ability. Skewness dispersion also delivers substantial economic gains in portfolio allocation. Its forecasting power is concentrated in months with monetary policy announcements, reflecting an information-based mechanism. The empirical evidence suggests that skewness dispersion captures the gradual incorporation of macro news into prices, which is driven by variation in aggregate risk and valuation adjustments.

Dissecting AI Trading: Behavioral Finance and Market Bubbles

S. Ouyang, P. Sui

University of Oxford, The Chinese University of Hong Kong, Shenzhen

We study how AI agents form expectations and trade in experimental asset markets. Using a simulated open-call auction populated by autonomous Large Language Model (LLM) agents, we document three main findings. First, AI agents exhibit classic behavioral patterns: a pronounced disposition effect and recency-weighted extrapolative beliefs. Second, these individual-level patterns aggregate into equilibrium dynamics that replicate classic experimental findings (Smith et al. (1988)), including the predictive power of excess demand for future prices and the positive relationship between disagreement and trading volume. Third, by analyzing the agents' reasoning text through a twenty-mechanism scoring framework, we show that targeted prompt interventions causally amplify or suppress specific behavioral mechanisms, significantly altering the magnitude of market bubbles.

AI "Errors"

W. Huang, A. J. Menkveld, S. Yu

Bank for International Settlements, Vrije Universiteit Amsterdam, Singapore Management University

When AIs are tasked with empirical research, how do their outcomes compare to those of humans? Are the distributions similar? We run an experiment where we let AI models repeat an experiment that was run with 164 human teams. Not surprisingly, distributions differ. The deeper question is: Why? We develop an approach that identifies which decisions on the analysis path drive these differences, which we define as "AI errors." The results show that AI concentrates on a narrow set of analysis paths, yielding markedly lower dispersion. For complex tasks, their estimates are systematically shifted relative to the human benchmark. Fork-level diagnostics and quantile regressions trace these shifts primarily to the choice of the statistical model, e.g., identifying a time trend by adding a stationary trend to a model in levels, or by computing relative changes and taking the average.

Mathematical and Computational Finance

Bridging the Reality Gap in Limit Order Book Simulation

P. Noble, M. Rosenbaum, S. Souilmi

Jump Trading, Université Paris Dauphine-PSL, École Polytechnique

We introduce a practical, interactive simulator of the limit order book for large-tick assets, designed to produce realistic execution, costs, and P&L. The book state is projected onto a tractable representation based on spread and volume imbalance, enabling robust estimation from market data. Event timing is calibrated to reproduce the fine-scale temporal structure of real markets, revealing a pronounced mode at exchange round-trip latency consistent with simultaneous reactions and latency races among participants. We further incorporate a feedback mechanism that accumulates signed trade flow through a power-law decay kernel, reproducing both concave market impact during execution and partial post-trade reversion. Across several stocks and strategy case studies, the simulator yields realistic behavior where profitability becomes highly sensitive to execution parameters. We present the approach as a practical recipe: project, estimate, validate, adapt, for building realistic limit order book simulations.

DORADO: Dynamic Optimization of R&D Options

Z. Chen, L. Kogan, A. W. Lo, Q. Xu, R. Zhang

Peking University, Massachusetts Institute of Technology

We propose a dynamic strategy to optimize a portfolio of research and development (R&D) projects with correlated binary outcomes, which we call dynamic optimization of R&D options (DORADO). DORADO utilizes project correlation as a knowledge discovery mechanism in order to exercise real options with the information learned from previous outcomes to dynamically adjust future R&D plans. We focus on a class of R&D projects, known as long shots, that are difficult to manage using standard approaches. We derive the DORADO-optimal strategy and its corresponding financial value for these portfolios under batch testing, block correlation, and risk-sensitive objectives. We further extend DORADO with a reinforcement learning policy with offline model estimation to accommodate R&D projects with heterogeneous characteristics at scale. DORADO generates an improved efficient frontier with a higher expected payoff and lower volatility than the comparable parallel development of these portfolios. Depending on investor preferences, the financial value of such an investment portfolio is maximized when the correlation between projects is either very low or very high, which reflects the tradeoff between the diversification gains at low correlations and the information value gained at high correlations. We provide an empirical example in the context of curative treatments for glioblastoma. Our results have important implications for managing highly correlated RD projects that cannot be effectively diversified.

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 members at the institute: Patrick Chang, Fayçal Drissi, Christopher Hyland, Jiawen Liang, and Kevin Schneider.

References

Ahrens, Achim, Victor Chernozhukov, Christian Hansen, Damian Kozbur, Mark Schaffer, and Thomas Wiemann. 2026. "An Introduction to Double/Debiased Machine Learning." *arXiv preprint arXiv:2504.08324*.

Babiak, Mykola, Jozef Barunik, and Josef Kurka. 2026. "Skewness Dispersion and Stock Market Returns." *arXiv preprint arXiv:2604.07870*.

Bykhovskaya, Anna, and Nour Meddahi. 2026. "Generalized Autoregressive Multivariate Models: From Binary to Poisson." *arXiv preprint arXiv:2604.14394*.

Chen, Zixi, Leonid Kogan, Andrew W Lo, Qingyang Xu, and Ruixun Zhang. 2026. "DORADO: Dynamic Optimization of R&D Options." *Available at SSRN 6293478*.

Christensen, Kim, Wenjing Liu, Zhi Liu, and Yoann Potiron. 2026. "The Realized Copula of Volatility." *arXiv preprint arXiv:2604.15811*.

Didisheim, Antoine, Bryan T Kelly, Mohammad Pourmohammadi, and Hanqing Tian. 2026. "The Inefficient Pricing of News." *NBER Working Paper No. 35093*.

Feng, Long, Huifang Ma, and Zhaojun Wang. 2026. "Testing Alpha in High-Dimensional Conditional Time-Varying Factor Models with Dependent Observations." *arXiv preprint arXiv:2604.13772*.

Ge, Shuyi, Shaoran Li, Oliver B Linton, and Wen Su. 2026. "Text as Priors." *Available at SSRN 6555359*.

Huang, Wenqian, Albert J Menkveld, and Shihao Yu. 2026. "AI "Errors"." *Available at SSRN 6408138*.

Noble, Patrick, Mathieu Rosenbaum, and Saad Souilmi. 2026. "Bridging the Reality Gap in Limit Order Book Simulation." *arXiv preprint arXiv:2603.24137*.

Ouyang, Shumiao, and Pengfei Sui. 2026. "Dissecting AI Trading: Behavioral Finance and Market Bubbles." *Available at SSRN 6589539*.

Souza, André, Christian T Brownlees, and Carlo Pavanello. 2026. "Which Factors Drive Downside Risk in the US Economy?" *Available at SSRN 6548438*.

Timmermann, Allan, and Luka Vulicevic. 2026. "Compute, Complexity, and the Scaling Laws of Return Predictability." *Available at SSRN 6105327*.

Yiu, Horace, Leandro Sánchez-Betancourt, Álvaro Cartea, and Gerardo Duran-Martin. 2026. "Doubly Outlier-Robust Online Infinite Hidden Markov Model." *arXiv preprint arXiv:2604.14322*.