VOL. 5, NO. 9

OMI Research Newsletter – June 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.*

Highlights

Key contributions in this month's edition include the following. Christensen, Goncalves and Perron (2026) show that even small misclassification errors in AI/ML-generated binary labels can substantially bias regression estimates, and propose a coupled-label bootstrap to restore valid inference. Bollerslev et al. (2026) develop candlestick-based spot correlation estimators using high-frequency open, close, high, and low prices, and document rapid changes in stock–bond correlations around FOMC announcements and tariff news. Cartea, Drissi and Saleh (2026) show that stricter regulation of tokenized securities reduces market liquidity and can make tokenization commercially non-viable, with the viability threshold depending on the asset's trading and compliance characteristics. Chen, Sialm and Xu (2026) show that AI-driven investing has grown mainly among hedge funds, generated early outperformance that faded after 2017, and appears more diverse than comparable non-AI strategies. Huang, Julliard and Shi (2026) argue that the seemingly dense factor zoo can be reduced to a few economically interpretable "factor identities", especially macro, market-level, and latent statistical pricing components. Doshi and Jacobs (2026) decompose CAPM beta into positive and negative cross-stock covariance components, and find that only the negative beta—interpretable as a market hedge—earns a robust and economically large risk premium. Choi, Lee and Kim (2026) develop a high-frequency, characteristic-based framework that separates diffusive risk, jump risk, and mispricing in stock returns, showing that these components arise from different firm characteristics and that the resulting intraday arbitrage signal earns strong risk-adjusted returns. Finally, Farboodi, Koh and Xia (2026) model how task-specific data fuels automation, showing that data spillovers can eventually automate the whole economy, but that the transition is slow, uneven, inefficient, and may generate explosive growth alongside stagnant wages.

Statistics, Econometrics & Machine Learning

Bootstrapping with AI/ML-generated labels

T. CHRISTENSEN, S. GONCALVES, B. PERRON
YALE UNIVERSITY, MCGILL UNIVERSITY, UNIVERSITÉ DE MONTRÉAL

AI/ML methods are increasingly used in economics to generate binary variables (or labels) via classification algorithms. When these generated variables are included as covariates in regressions, even small misclassification errors can induce large biases in OLS estimators and invalidate standard inference. We study whether the bootstrap can correct this bias and deliver valid inference. We first show that a seemingly natural fixed-label bootstrap, which generates data using estimated labels but relies on a corrupted version in estimation, is generally invalid unless a strong independence condition between the latent true labels and other covariates holds. We then propose a coupled-label bootstrap that jointly resamples the true and imputed labels, and show it is valid without this condition. Two finite-sample adjustments further improve coverage: a variance correction for uncertainty in estimated misclassification rates and a Hessian rotation for near-singular designs. We illustrate the methods in simulations and apply them to investigate the relationship between wages and remote work status.

Estimating Correlations and Reading the News by Candlesticks

T. Bollerslev, J. Li, Q. Li, Y. Li
Duke University, Singapore Management University, University of Hong Kong, University of Manchester

We develop a new class of nonparametric spot correlation estimators based on high-frequency candlestick data comprising open, close, high, and low prices over short intraday time intervals. The new estimators augment the conventional return-based estimator with a signed asymmetry feature. We establish consistency and mixed normality in a traditional large-estimation-window in-fill asymptotic setting. We also develop a fixed-k asymptotic framework and confidence intervals that achieve near-nominal finite-sample coverage with very few candlesticks. Simulations further demonstrate substantial mean-squared-error reductions over return-based estimators and robustness to market microstructure noise. Relying on one-minute candlesticks, we find that FOMC announce-ments systematically raise stock-bond correlations, consistent with a discount-rate channel interpretation of the news, while President Trump's tariff-related posts affect more dispersed changes in the correlations. The empirical results highlight the value of the new candlestick-based spot estimators for capturing rapid changes in asset comovements around economic news and illuminating the economic transmission mechanisms at work.

On Lead-Lag Estimation of Non-Synchronously Observed Point Processes

T. Shiotani, T. Hayashi, Y. Koike
University of Tokyo, Keio University

This paper introduces a new theoretical framework for analyzing lead-lag relationships between point processes, with a special focus on applications to high-frequency financial data. In particular, we are interested in lead-lag relationships between two sequences of order arrival timestamps. The seminal work of Dobrev and Schaumburg proposed model-free measures of cross-market trading activity based on cross-counts of timestamps. While their method is known to yield reliable results, it faces limitations because its original formulation inherently relies on discrete-time observations, an issue we address in this study. Specifically, we formulate the problem of estimating lead-lag relationships in two point processes as that of estimating the shape of the cross-pair correlation function (CPCF) of a bivariate stationary point process, a quantity well-studied in the neuroscience and spatial statistics literature. Within this framework, the prevailing lead-lag time is defined as the location of the CPCF's sharpest peak. Under this interpretation, the peak location in Dobrev and Schaumburg's cross-market activity measure can be viewed as an estimator of the lead-lag time in the aforementioned sense. We further propose an alternative lead-lag time estimator based on kernel density estimation and show that it possesses desirable theoretical properties and delivers superior numerical performance. Empirical evidence from high-frequency financial data demonstrates the effectiveness of our proposed method.

Group Network Multivariate GARCH

J. Chen, W. Ma, G. Xu
University of Sussex, University of Pennsylvania, University of Miami

Traditional multivariate generalized autoregressive conditional heteroskedasticity (GARCH) models often become infeasible in high-dimensional settings because of the curse of dimensionality. We propose a group-network multivariate GARCH model in which the propagation of past return shocks and past conditional variances across assets is governed by a latent group structure and an observed network adjacency matrix. By imposing network- and group-based restrictions, the proposed model substantially reduces the parameter dimension, thereby facilitating estimation in large cross-sections. We establish the large-sample properties of an estimator that jointly estimates the model parameters and latent group memberships via an iterative optimization algorithm. Monte Carlo simulations support the theoretical results. An empirical application to S&P 100 constituents over 2015-2022 shows that the proposed model delivers economically meaningful gains in out-of-sample portfolio performance compared to standard multivariate GARCH specifications.

Finance and Economics

The Economics of Regulating Tokenized Securities

Á. Cartea, F. Drissi, F. Saleh
Oxford-Man Institute, University of Oxford, Southern University of Florida

We develop an economic model to study the effect of regulation on tokenized securities markets. Tokenizing a security does not exempt it from securities laws, so in our framework token issuers must choose a compliance policy that restricts trading. We characterize the resulting trade-off between regulatory stringency and financial market outcomes, and we derive three main results. First, tighter regulation reduces the liquidity of the tokenized market. Second, tokenization is not commercially viable for sufficiently strict regulation. Third, the tightest level of regulation that tokenization can withstand depends on the characteristics of the tokenized asset.

The Growth and Performance of Artificial Intelligence in Asset Management

S. Chen, C. Sialm, D. X. Xu
University of Melbourne, University of Texas at Austin, Southern Methodist University

We examine the growth and performance of AI-driven investing. Using investment advisers' regulatory disclosures, labor market data, and fund strategy descriptions, we document that AI-driven investing has grown steadily since the early 2010s and is concentrated among hedge funds. AI hedge funds outperformed non-AI hedge funds in the early years, but this outperformance declined over time, even among early adopters. Contrary to concerns about AI-driven strategy homogeneity, AI hedge funds exhibit lower return comovement than non-AI peers. Our findings highlight both the alpha generating potential and the limitations of AI as a source of investment performance.

Factor Identity

J. Huang, C. Julliard, R. Shi
The University of Hong Kong, London School of Economics and Political Science, University of Colorado Boulder

The empirical dominance of dense factor models for pricing the cross-section of expected returns has come at a cost: the economic narrative of priced risk has disappeared. We introduce factor identities-groups of factors conveying the same pricing information despite modest correlations-and a Bayesian method to recover them, restoring legibility to the dense SDF. Only a handful emerge: a macroeconomic identity revealed through long-horizon growth in output, consumption, and industrial production; a market identity pinning down the level of risk premia; and a few characteristic-based identities that are fully subsumed by latent factors. The factor zoo is dense in animals but sparse in species.

AlphaGlass: Interpretable Characteristic-Based Portfolio Choice

S. Bell, A. Kakhbod, M. Lettau, A. Nazemi
Karlsruhe Institute of Technology, University of California, Berkeley, NBER

We propose AlphaGlass, an inherently interpretable machine-learning framework for constructing portfolios that directly optimize investment objectives. AlphaGlass maps stock characteristics into additive signals with sparse interactions and converts these signals into long-short portfolios through a differentiable rank-and-mask layer. This end-to-end design allows the model to optimize objectives such as the Sharpe ratio or mean-variance utility while keeping portfolio weights interpretable and traceable to specific characteristics and interactions. We show theoretically that in-sample objective maximization consistently estimates the population objective and that the differentiable rank-and-mask layer is a faithful smooth proxy for the corresponding conventional long-short quantile portfolio. In U.S. equities, AlphaGlass delivers strong out-of-sample performance and reveals economically interpretable drivers of long and short positions.

A Cross-Sectional Decomposition of Firms' Market Betas

H. Doshi, K. Jacobs
University of Houston

We introduce a new cross-sectional model that generalizes the CAPM. Using the linear expansion property of covariance, we decompose firms' CAPM betas into negative betas, which capture the negative covariances with other firms in the value-weighted market portfolio and provide a hedge against the market, and positive betas. These two betas sum to market beta, but only negative beta commands statistically and economically significant risk premia. This finding is robust against existing factors and characteristics, highlighting a distinct dimension of systematic risk and providing a partial explanation for several existing anomalies. Negative beta also differs from downside beta and semibetas.

Measuring Bubbles via Put-Call Disparity: A Model-Free Approach

R. A. Jarrow, S. Kwok
Cornell University, University of Sydney

This paper introduces simple, model-free lower and upper bounds for measuring the size of asset price bubbles. Assuming only that the market satisfies no-free-lunch-with-vanishing-risk and that all trading strategies are admissible, our framework avoids restrictive parametric models and the no-dominance assumption. We demonstrate that put-call disparity provides an observable lower bound, and is economically justified by short-sale constraints. Additionally, the lowest price of an out-of-the-money (OTM) call option determines the upper bound. To ensure empirical reliability, we modify these bounds using data-driven regularization and bootstrap methods to disentangle genuine bubble signals from market microstructure noise and to reduce reliance on thinly traded deep OTM options. Using S&P 500 index option prices from 1996 to 2025, we document a sustained bubble during the COVID-19 era and capture market exuberance preceding the 2000 dot-com and 2008 financial crashes. In addition, the empirical study suggests that the market violates no-dominance and is incomplete.

Characteristic-Based Risk Decomposition and Mispricing in High-Frequency Return Panels

Y. Choi, S. S. Lee, S. Kim

Xavier University, Georgia Institute of Technology, Korea Advanced Institute of Science and Technology (KAIST)

We propose a novel framework that integrates firm characteristics into the estimation of systematic diffusive and jump risks using high-frequency return panels, allowing us to identify mispricing that is orthogonal to the decomposed risks. Applying the framework to more than 6,600 U.S. equities from 2000–2021, we estimate characteristic-based diffusive and jump factor betas and identify a mispricing component. We show that diffusive risk, jump risk, and mispricing are associated with distinct clusters of firm characteristics, revealing differences in their economic sources and time-series variation. The resulting intraday arbitrage portfolio delivers substantial risk-adjusted returns. We further show that the characteristic-based diffusive and jump factors command significant risk premia, with signed jump risks playing economically distinct roles.

R-Squareds in the Cross Section and Time Series

P. Balduzzi, T. Bonafini, C. Robotti, R. Savona

Boston College, Free University of Bozen-Bolzano, University of Warwick, University of Brescia

We exploit the link between systematic volatility and betas in linear factor models to show that the R-squared of a cross-sectional (CS) second-pass regression of test asset risk premia on factor loadings can be factorized into two components. The first component captures the factor model's average time-series (TS) fit across test assets, while the second component reflects the maximum risk-return trade-off of the model's optimal factor portfolio (OFP) relative to that of the test assets.

Our factorization explains why adding systematic factors to a model may reduce the TS fit: the new OFP hedges away systematic risk and explains less of the test assets' volatility. Our decomposition also explains why weak factors, while not systematic, can improve pricing: they have higher information ratios than the factors in existing models.

For example, the improvement in pricing performance going from the capital asset pricing model to three-, five-, and six-factor Fama-French specifications is mainly due to an increase in the relative risk-return trade-off of the factors of the larger models, and not to an improvement in the average TS fit. Furthermore, we employ our novel decomposition to show that factors other than the market enter the 'factor zoo' mainly because they are profitable in sample, not because they are pervasive. The market, on the other hand, dominates all other factors in terms of its pervasiveness. Finally, our out-of-sample results provide an interpretation for the longevity of the CAPM and for the lack of robustness and associated skepticism regarding all the empirically-motivated pricing factors that have been proposed in literature.

The Intramonth Momentum Cycle

D. Nathan, M. Suominen, J. Tasa

Hong Kong Polytechnic University, Aalto University

U.S. equity momentum returns concentrate in six trading days each month, ending four trading days before month-end. We trace this concentration to predictable month-end cash demand: investors sell loser stocks for liquidity. Over 1980–2025, $1 invested in a value-weighted Winners-Minus-Losers portfolio grows to $18.78 in these six days, versus $2.37 otherwise. The effect is driven by bottom-decile losers underperforming the market, while winners show no comparable concentration, and the same loser-driven pattern appears across 19 developed markets. We introduce the concept of asset dispensability to explain why the selling falls on losers: stocks with embedded tax losses, low foregone dividend income, and salient underperformance are the natural candidates to liquidate when cash is needed. A measure capturing dispensability directly predicts the concentration better than past returns themselves, and sorting on it produces a larger six-day premium. Settlement-cycle reforms validate the timing: U.S. $(T+2)$-to-$(T+1)$ and synchronized European $(T+3)$-to-$(T+2)$ transitions each move loser-selling pressure one trading day closer to month-end. Momentum largely reflects predictable liquidity demand.

The Value of Information: A Puzzle

O. Kadan, A. Manela

Arizona State University, Washington University in St. Louis

We show that under mild assumptions, the total value of information to informed traders in the market can be measured by the covariance between price changes and order flow. This covariance captures noise trader losses, which equal informed trader gains when market making is competitive. We estimate the value of information using high frequency data on US equities at about $3.5 million per year for the average stock. The aggregate value of information is about 0.04% of market cap, which is considerably lower than the 0.67% in fees investors pay each year searching for superior returns (French 2008). We discuss potential resolutions for these puzzling findings.

Data-Driven Automation

M. Farboodi, A. Koh, A. Xia

MIT Sloan School of Management

We build a dynamic model of data-driven automation in which data (i) is heterogeneous and task-specific; (ii)

accumulates endogenously as a byproduct of economic activity; and (iii) exhibits spillovers such that data generated by one task can augment the productivity of another. Along the transition path of automation, data plays a dual role in simultaneously augmenting the productivity of already-automated tasks and expanding the automation frontier. We derive tight conditions for the economy to be partially versus fully automated in the long-run. In the latter case, automation exhibits rich short-run dynamics that depend on the pattern of data spillovers but is always slow in the long-run: the share of tasks produced by labor decays asymptotically as a power law in time. We show that the economy is generically inefficient and analyze how a planner optimally tilts the direction of data accumulation. With endogenous capital accumulation, data-driven automation generates explosive growth but stagnant long-run wages.

Mathematical and Computational Finance

Optimal Control of the Ethena Yield-Bearing Stablecoin

M. Lorig
University of Washington

We formulate and solve stochastic control problems that model the core yield-generating strategy of the Ethena protocol, a decentralized finance (DeFi) stablecoin that earns yield by combining a long position in staked Ethereum (stETH) with an equal-sized short position in ETH perpetual futures. The combined position is delta-neutral with respect to the ETH spot price, yet earns carry from two sources: staking rewards on the stETH leg, and funding-rate payments received from long perpetual holders when the perpetual trades at a premium to spot. A key feature of our model is that the control – the rate of simultaneously buying stETH and shorting the perpetual – exerts two distinct types of price impact. *Permanent* impact shifts the mid-market prices of both legs, compressing the basis and permanently eroding future funding income. *Temporary* impact reflects execution slippage on each leg. We study both an infinite-horizon discounted problem and a finite-horizon problem in which the protocol maximizes total wealth up to a fixed date T, subject to a terminal cost for liquidating any remaining position. In both cases the optimal control is obtained explicitly.

Enhancing a Risk Model by Adding Transient Statistical Factors

A. E. Tzikas, E. J. Candès, T. Hastie, S. P. Boyd, M. J. Kochenderfer, R. N. Kahn
Stanford University, BlackRock

Estimating the covariance of asset returns, i.e., the risk model, is a key component of financial portfolio construction and evaluation. Most risk modeling approaches produce a factor model that decomposes the asset variability into two components: the first attributed to a small number of factors that are common among the assets and the second attributed to the idiosyncratic behavior of each asset. Third-party providers typically provide risk models to investors, and while these models are typically of high quality, they may fail to capture important information, e.g., changing market regimes and transient factors. To overcome these limitations, we propose a systematic method based on maximum likelihood estimation to enhance an existing factor model by both refining the given model and adding new statistical factors. Our approach relies only on the observed sequence of realized returns and on the choice of two hyperparameters: the number of additional factors and the half-life parameter that determines the weights assigned to returns in the log-likelihood objective. Importantly, our methodology applies to the situation where asset returns may be missing, making it suitable for typical equity datasets. We demonstrate our approach on the Barra short-term US risk model, a high-quality risk model used in practice, for a universe of US high-capitalization equities. We show that the proposed extension captures structure in the returns that is missed by the original model.

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

Balduzzi, Pierluigi, Tommaso Bonafini, Cesare Robotti, and Roberto Savona. 2026. "R-Squareds in the Cross Section and Time Series." *Available at SSRN 6426158*.

Bell, Sebastian, Ali Kakhbod, Martin Lettau, and Abdolreza Nazemi. 2026. "AlphaGlass: Interpretable Characteristic-Based Portfolio Choice." *NBER Working Paper w35186*.

Bollerslev, Tim, Jia Li, Qiyuan Li, and Yifan Li. 2026. "Estimating Correlations and Reading the News by Candlesticks." *Working Paper*.

Cartea, Álvaro, Fayçal Drissi, and Fahad Saleh. 2026. "The Economics of Regulating Tokenized Securities." *Available at SSRN 6942338*.

Chen, Jian, Weidong Ma, and Ganggang Xu. 2026. "Group Network Multivariate GARCH." *Available at SSRN 6760338*.

Chen, Shuang, Clemens Sialm, and David X Xu. 2026. "The Growth and Performance of Artificial Intelligence in Asset Management." *Available at SSRN 5638612*.

Choi, Youngmin, Suzanne S Lee, and Soohun Kim. 2026. "Characteristics-based Risk Decomposition and Mispricing in High-frequency Return Panels." *Available at SSRN 5632370*.

Christensen, Timothy, Silvia Goncalves, and Benoit Perron. 2026. "Bootstrapping with AI/ML-generated labels." *arXiv preprint arXiv:2604.23770*.

Doshi, Hitesh, and Kris Jacobs. 2026. "A Cross-Sectional Decomposition of Firms' Market Betas." *Available at SSRN 5279556*.

Farboodi, Maryam, Andrew Koh, and Anchi Xia. 2026. "Data-Driven Automation." *arXiv preprint arXiv:2606.10127*.

Huang, Jiantao, Christian Julliard, and Ran Shi. 2026. "Factor Identity." *Available at SSRN 6404361*.

Jarrow, Robert A, and Simon Kwok. 2026. "Measuring Bubbles via Put-Call Disparity: A Model-Free Approach." *Available at SSRN 6754305*.

Kadan, Ohad, and Asaf Manela. 2026. "The Value of Information: A Puzzle." *arXiv preprint arXiv:2605.11180*.

Lorig, Matthew. 2026. "Optimal Control of the Ethena Yield-Bearing Stablecoin." *arXiv preprint arXiv:2605.11263*.

Nathan, Daniel, and Matti Suominen. 2026. "The Intramonth Momentum Cycle." *Available at SSRN 6426026*.

Shiotani, Takaaki, Takaki Hayashi, and Yuta Koike. 2026. "On Lead-Lag Estimation of Non-Synchronously Observed Point Processes." *arXiv preprint arXiv:2601.01871*.

Tzikas, Alexandros E, Emmanuel J Candes, Trevor Hastie, Stephen P Boyd, Mykel J Kochenderfer, and Ronald N Kahn. 2026. "Enhancing a Risk Model by Adding Transient Statistical Factors." *arXiv preprint arXiv:2605.12977*.