OMI Research Newsletter – January 2024



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

The OMI newsletter team wishes all our readers a happy holiday season. We are looking forward to another year of great research.

Research focus. Key contributions in this month's edition include works from the OMI, Duke University, University of Washington, and Queen Mary University. Cartea et al. (2023) propose a new method based on a a pairwise measure of dissimilarity to forecast the realised covariance of multidimensional time series. Suder, Xu and Dunson (2023) review the concepts of Transfer Learning through a Bayesian lens. Ohnishi et al. (2023) generalises the Bellman equation to the space of trajectories using Signatures. Albers et al. (2023) present findings of a large-scale trading experiment in cryptocurrency exchanges and show that volatility, latency, and liquidity impact execution outcomes.

Machine Learning & Statistics

A Similarity-based Approach to Covariance Forecasting

Á. CARTEA, M. CUCURINGU, M. JENNINGS, C. ZHANG

OXFORD-MAN INSTITUTE

We introduce a novel method to forecast the realised covariances of multidimensional time series. The central innovation of our method is a two-step procedure which first filters the training data by relevance before using non-parametric or linear approaches to make predictions. The relevance filter uses the recent histories of both the test input and each individual training input to generate a pairwise measure of dissimilarity between the two points. Highly dissimilar training data which do not share the observed dynamics of the test input are discarded from the regression procedure, reducing the complexity of the forecasting task. The second step uses simple non-parametric or linear methods to perform regression with the relevant data from the first step. Our method produces fast, interpretable predictions which achieve state-of-the-art performance, and functions well in limited-data availability and high dimensional settings such as large covariance matrices because it does not depend on highly parameterised models. Furthermore, we propose a dynamic ensembling scheme to combine our method with existing models. The scheme uses information about the recent performance of forecasts from different models to produce a weighted average forecast which heavily discounts models ill-suited to the current market dynamics. We test our methods on realised covariance data of U.S. equities and find significant improvements over widely used benchmarks.

Bayesian Transfer Learning

P. M. SUDER, J. XU, D. B. DUNSON

DUKE UNIVERSITY

Transfer learning is a burgeoning concept in statistical machine learning that seeks to improve inference and/or predictive accuracy on a domain of interest by leveraging data from related domains. While the term "transfer learning" has garnered much recent interest, its foundational principles have existed for years under various guises. Prior literature reviews in computer science and electrical engineering have sought to bring these ideas into focus, primarily surveying general methodologies and works from these disciplines.

This article highlights Bayesian approaches to transfer learning, which have received relatively limited attention despite their innate compatibility with the notion of drawing upon prior knowledge to guide new learning tasks. Our survey encompasses a wide range of Bayesian transfer learning frameworks applicable to a variety of practical settings. We discuss how these methods address the problem of finding the optimal information to transfer between domains, which is a central question in transfer learning. We illustrate the utility of Bayesian transfer learning methods via a simulation study where we compare performance against frequentist competitors.

Implicit Maximum a Posteriori Filtering via Adaptive Optimization

G. M. Bencomo, J. C. Snell, T. L. Griffiths
Princeton University

Bayesian filtering approximates the true underlying behavior of a time-varying system by inverting an explicit generative model to convert noisy measurements into state estimates. This process typically requires either storage, inversion, and multiplication of large matrices or Monte Carlo estimation, neither of which are practical in high-dimensional state spaces such as the weight spaces of artificial neural networks. Here, we frame the standard Bayesian filtering problem as optimization over a time-varying objective. Instead of maintaining matrices for the filtering equations or simulating particles, we specify an optimizer that defines the Bayesian filter implicitly. In the linear-Gaussian setting, we show that every Kalman filter has an equivalent formulation using K steps of gradient descent. In the nonlinear setting, our experiments demonstrate that our framework results in filters that are effective, robust, and scalable to high-dimensional systems, comparing well against the standard toolbox of Bayesian filtering solutions. We suggest that it is easier to fine-tune an optimizer than it is to specify the correct filtering equations, making our framework an attractive option for high-dimensional filtering problems.

Mamba: Linear-Time Sequence Modeling with Selective State Spaces

A. Gu, T. Dao
Carnegie Mellon University, Princeton University

Foundation models, now powering most of the exciting applications in deep learning, are almost universally based on the Transformer architecture and its core attention module. Many subquadratic-time architectures such as linear attention, gated convolution and recurrent models, and structured state space models (SSMs) have been developed to address Transformers' computational inefficiency on long sequences, but they have not performed as well as attention on important modalities such as language. We identify that a key weakness of such models is their inability to perform content-based reasoning, and make several improvements. First, simply letting the SSM parameters be functions of the input addresses their weakness with discrete modalities, allowing the model to selectively propagate or forget information along the sequence length dimension depending on the current token. Second, even though this change prevents the use of efficient convolutions, we design a hardware-aware parallel algorithm in recurrent mode. We integrate these selective SSMs into a simplified end-to-end neural network architecture without attention or even MLP blocks (Mamba). Mamba enjoys fast inference (5× higher throughput than Transformers) and linear scaling in sequence length, and its performance improves on real data up to million-length sequences. As a general sequence model backbone, Mamba achieves state-of-the-art performance across several modalities such as language, audio, and genomics. On language modeling, our Mamba-3B model outperforms Transformers of the same size and matches Transformers twice its size, both in pretraining and downstream evaluation.

Towards Sobolev Pruning

N. Kichler, S. Afghan, U. Naumann
RWTH Aachen University

The increasing use of stochastic models for describing complex phenomena warrants surrogate models that capture the reference model characteristics at a fraction of the computational cost, foregoing potentially expensive Monte Carlo simulation. The predominant approach of fitting a large neural network and then pruning it to a reduced size has commonly neglected shortcomings. The produced surrogate models often will not capture the sensitivities and uncertainties inherent in the original model. In particular, (higher-order) derivative information of such surrogates could differ drastically. Given a large enough network, we expect this derivative information to match. However, the pruned model will almost certainly not share this behavior. In this paper, we propose to find surrogate models by using sensitivity information throughout the learning and pruning process. We build on work using Interval Adjoint Significance Analysis for pruning and combine it with the recent advancements in Sobolev Training to accurately model the original sensitivity information in the pruned neural network based surrogate model. We experimentally underpin the method on an example of pricing a multidimensional Basket option modelled through a stochastic differential equation with Brownian motion. The proposed method is, however, not limited to the domain of quantitative finance, which was chosen as a case study for intuitive interpretations of the sensitivities. It serves as a foundation for building further

surrogate modelling techniques considering sensitivity information.

Can a Transformer Represent a Kalman Filter?

G. Goel, P. Bartlett
Simons Institute, UC Berkeley

Transformers are a class of autoregressive deep learning architectures which have recently achieved state-of-the-art performance in various vision, language, and robotics tasks. We revisit the problem of Kalman Filtering in linear dynamical systems and show that Transformers can approximate the Kalman Filter in a strong sense. Specifically, for any observable LTI system we construct an explicit causally-masked Transformer which implements the Kalman Filter, up to a small additive error which is bounded uniformly in time; we call our construction the Transformer Filter. Our construction is based on a two-step reduction. We first show that a softmax self-attention block can exactly represent a certain Gaussian kernel smoothing estimator. We then show that this estimator closely approximates the Kalman Filter. We also investigate how the Transformer Filter can be used for measurement-feedback control and prove that the resulting nonlinear controllers closely approximate the performance of standard optimal control policies such as the LQG controller.

Castor: Causal Temporal Regime Structure Learning

A. Rahmani, P. Frossard
EPFL

The task of uncovering causal relationships among multivariate time series data stands as an essential and challenging objective that cuts across a broad array of disciplines ranging from climate science to healthcare. Such data entails linear or non-linear relationships, and usually follow multiple a priori unknown regimes. Existing causal discovery methods can infer summary causal graphs from heterogeneous data with known regimes, but they fall short in comprehensively learning both regimes and the corresponding causal graph. In this paper, we introduce CASTOR, a novel framework designed to learn causal relationships in heterogeneous time series data composed of various regimes, each governed by a distinct causal graph. Through the maximization of a score function via the EM algorithm, CASTOR infers the number of regimes and learns linear or non-linear causal relationships in each regime. We demonstrate the robust convergence properties of CASTOR, specifically highlighting its proficiency in accurately identifying unique regimes. Empirical evidence, garnered from exhaustive synthetic experiments and two real-world benchmarks, confirm CASTOR's superior performance in causal discovery compared to baseline methods. By learning a full temporal causal graph for each regime, CASTOR establishes itself as a distinctly interpretable method for causal discovery in heterogeneous time series.

Sparse Training of Discrete Diffusion Models for Graph Generation

Y. Qin, C. Vignac, P. Frossard
EPFL

Generative models for graphs often encounter scalability challenges due to the inherent need to predict interactions for every node pair. Despite the sparsity often exhibited by real-world graphs, the unpredictable sparsity patterns of their adjacency matrices, stemming from their unordered nature, leads to quadratic computational complexity. In this work, we introduce SparseDiff, a denoising diffusion model for graph generation that is able to exploit sparsity during its training phase. At the core of SparseDiff is a message-passing neural network tailored to predict only a subset of edges during each forward pass. When combined with a sparsity-preserving noise model, this model can efficiently work with edge lists representations of graphs, paving the way for scalability to much larger structures. During the sampling phase, SparseDiff iteratively populates the adjacency matrix from its prior state, ensuring prediction of the full graph while controlling memory utilization. Experimental results show that SparseDiff simultaneously matches state-of-the-art in generation performance on both small and large graphs, highlighting the versatility of our method.

Signatures Meet Dynamic Programming: Generalizing Bellman Equations for Trajectory Following

M. Ohnishi, I. Akinola, J. Xu, A. Mandlekar, F. Ramos
University of Washington

Path signatures have been proposed as a powerful representation of paths that efficiently captures the path's analytic and geometric characteristics, having useful algebraic properties including fast concatenation of paths through tensor products. Signatures have recently been widely adopted in machine learning problems for time series analysis. In this work we establish connections between value functions typically used in optimal control and intriguing properties of path signatures. These connections motivate our novel control framework with signature transforms that efficiently generalizes the Bellman equation to the space of trajectories. We analyze the properties and advantages of the framework, termed signature control. In particular, we demonstrate that (i) it can naturally deal with varying/adaptive time steps; (ii) it propagates higher-level information more efficiently than value function updates; (iii) it is robust

to dynamical system misspecification over long rollouts. As a specific case of our framework, we devise a model predictive control method for path tracking. This method generalizes integral control, being suitable for problems with unknown disturbances. The proposed algorithms are tested in simulation, with differentiable physics models including typical control and robotics tasks such as point-mass, curve following for an ant model, and a robotic manipulator.

Finance and Economics

The Good, the Bad, and Latency: Exploratory Trading on Bybit and Binance

J. Albers, M. Cucuringu, S. Howison, A. Y. Shestopaloff

Oxford-Man Institute, University of Oxford, Queen Mary University

We present the findings of a large-scale live trading experiment involving the placement of millions of market orders sent at a high frequency on two cryptocurrency exchanges, Bybit and Binance. We analyze the execution outcomes of these orders in comparison to the expected outcome based on the most recent snapshot of the Limit Order Book (LOB) at the time of order submission for two execution modes: one using market orders and the second using marketable limit orders aiming at the best price. Discrepancies between the actual and expected outcomes are due to intermittent LOB updates during a time span resulting from delays on the exchange, delays on the trader's end, or communication delays between the trader and the exchange. We show these discrepancies are strongly correlated with market factors such as volatility, latency, and LOB liquidity. Notably, we find a consistent disadvantage to the trader, pointing to an adverse selection effect for taker orders: profitable orders (as measured by short-term future PnL returns) tend to achieve worse-than-expected outcomes, while unprofitable orders typically achieve their expected (adverse) outcomes. In the case of market orders, this translates to a worsening of fill prices, while marketable limit orders suffer from a substantial probability of failing-to-fill-immediately. Quantitative researchers who fail to take these effects into account face the familiar litany of underperforming in a live trading environment relative to stellar backtests. To address this concern, we propose parsimonious models to estimate an order's probability of failing-to-fill-immediately (in case of a marketable limit order) and the worsening of its fill price (in case of a market order), allowing for greater accuracy when carrying out backtests and minimizing the discrepancy between backtest and realized live PnL.

Algorithmic Persuasion Through Simulation: Information Design in the Age of Generative AI

K. Harris, N. Immorlica, B. Lucier, A. Slivkins

Carnegie Mellon University

How can an informed sender persuade a receiver, having only limited information about the receiver's beliefs? Motivated by research showing generative AI can simulate economic agents, we initiate the study of information design with an oracle. We assume the sender can learn more about the receiver by querying this oracle, e.g., by simulating the receiver's behavior. Aside from AI motivations such as general-purpose Large Language Models (LLMs) and problem-specific machine learning models, alternate motivations include customer surveys and querying a small pool of live users. Specifically, we study Bayesian Persuasion where the sender has a second-order prior over the receiver's beliefs. After a fixed number of queries to an oracle to refine this prior, the sender commits to an information structure. Upon receiving the message, the receiver takes a payoff-relevant action maximizing her expected utility given her posterior beliefs. We design polynomial-time querying algorithms that optimize the sender's expected utility in this Bayesian Persuasion game. As a technical contribution, we show that queries form partitions of the space of receiver beliefs that can be used to quantify the sender's knowledge.

The Corporate Bond Factor Zoo

A. Dickerson, C. Julliard, P. Mueller

UNSW Business School, LSE, Warwick Business School

Analyzing 563 trillion possible models, we find that the majority of tradable factors designed to price bond markets are unlikely sources of priced risk, and only one novel tradable bond factor, capturing the bond post-earnings announcement drift, should be included in the stochastic discount factor (SDF) with very high probability. Nevertheless, the SDF is dense in the space of observable factors, with both nontradable and equity-based ones being salient for pricing corporate bonds, and a Bayesian model averaging–SDF explains corporate risk premia better than all existing models, both in- and out-of-sample, and captures business cycle and market crash risks.

Insider trading in discrete time Kyle games

C. Kühn, C. Lorenz

Goethe University Frankfurt

We present a discrete time version of Kyle's (1985) classic model of insider trading. The model has three kinds of traders: an insider, random noise traders, and a market maker. The insider aims to exploit her informational advantage and maximise expected profits while the market maker observes the total order flow and sets prices accordingly. First, we show how the multi-

period model with finitely many pure strategies can be reduced to a (static) social system in the sense of Debreu (1952) and prove the existence of a sequential Kyle equilibrium, following Kreps and Wilson (1982). This requires no probabilistic restrictions on the true value, the insider's dynamic information, and the noise trader's actions. In the single-period model we establish bounds for the insider's strategy in equilibrium. Finally, we prove the existence of an equilibrium for the game with a continuum of actions, by considering an approximating sequence of games with finitely many actions. Because of the lack of compactness of the set of measurable price functions, standard infinite-dimensional fixed point theorems are not applicable.

Mathematical Finance

Factor Risk Budgeting and Beyond

A. R. Cetingoz, O. Guéant
Université Paris 1 Panthéon-Sorbonne

Portfolio optimization methods have evolved significantly since Markowitz introduced the mean-variance framework in 1952. While the theoretical appeal of this approach is undeniable, its practical implementation poses important challenges, primarily revolving around the intricate task of estimating expected returns. As a result, practitioners and scholars have explored alternative methods that prioritize risk management and diversification. One such approach is Risk Budgeting, where portfolio risk is allocated among assets according to predefined risk budgets. The effectiveness of Risk Budgeting in achieving true diversification can, however, be questioned, given that asset returns are often influenced by a small number of risk factors. From this perspective, one question arises: is it possible to allocate risk at the factor level using the Risk Budgeting approach? This paper introduces a comprehensive framework to address this question by introducing risk measures directly associated with risk factor exposures and demonstrating the desirable mathematical properties of these risk measures, making them suitable for optimization. We also propose a framework to find the portfolio that effectively balances the risk contributions from both assets and factors. Leveraging standard stochastic algorithms, our framework enables the use of a wide range of risk measures.

Limit Order Book Dynamics and Order Size Modelling Using Compound Hawkes Process

K. Jain, N. Firoozye, J. Kochems, P. Treleaven
affiliation

Hawkes Process has been used to model Limit Order Book (LOB) dynamics in several ways in the literature however the focus has been limited to capturing the inter-event times while the order size is usually assumed to be constant. We propose a novel methodology of using Compound Hawkes Process for the LOB where each event has an order size sampled from a calibrated distribution. The process is formulated in a novel way such that the spread of the process always remains positive. Further, we condition the model parameters on time of day to support empirical observations. We make use of an enhanced non-parametric method to calibrate the Hawkes kernels and allow for inhibitory cross-excitation kernels. We showcase the results and quality of fits for an equity stock's LOB in the NASDAQ exchange.

Optimal investment with a noisy signal of future stock prices

P. Bank, Y. Dolinsky
TU Berlin

We consider an investor who is dynamically informed about the future evolution of one of the independent Brownian motions driving a stock's price fluctuations. With linear temporary price impact the resulting optimal investment problem with exponential utility turns out to be not only well posed, but it even allows for a closed-form solution. We describe this solution and the resulting problem value for this stochastic control problem with partial observation by solving its convex-analytic dual problem.

Acknowledgement

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

References

Albers, Jakob, Mihai Cucuringu, Sam Howison, and Alexander Y. Shestopaloff. 2023. "The Good, the Bad, and Latency: Exploratory Trading on Bybit and Binance." *Available at SSRN 4677989*.

Bank, Peter, and Yan Dolinsky. 2023. "Optimal investment with a noisy signal of future stock prices." *arXiv preprint arXiv:2302.10485*.

Bencomo, Gianluca M, Jake C Snell, and Thomas L Griffiths. 2023. "Implicit Maximum a Posteriori Filtering via Adaptive Optimization." *arXiv preprint arXiv:2311.10580*.

Cartea, Álvaro, Mihai Cucuringu, Mark Jennings, and Chao Zhang. 2023. "A Similarity-based Approach to Covariance Forecasting." *Available at SSRN 4652980*.

Cetingoz, Adil Rengim, and Olivier Guéant. 2023. "Factor Risk Budgeting and Beyond." *arXiv preprint arXiv:2312.11132*.

Dickerson, Alexander, Christian Julliard, and Philippe Mueller. 2023. "The Corporate Bond Factor Zoo." *Available at SSRN*.

Goel, Gautam, and Peter Bartlett. 2023. "Can a Transformer Represent a Kalman Filter?" *arXiv preprint arXiv:2312.06937*.

Gu, Albert, and Tri Dao. 2023. "Mamba: Linear-Time Sequence Modeling with Selective State Spaces." *arXiv preprint arXiv:2312.00752*.

Harris, Keegan, Nicole Immorlica, Brendan Lucier, and Aleksandrs Slivkins. 2023. "Algorithmic Persuasion Through Simulation: Information Design in the Age of Generative AI." *arXiv preprint arXiv:2311.18138*.

Jain, Konark, Nick Firoozye, Jonathan Kochems, and Philip Treleaven. 2023. "Limit Order Book Dynamics and Order Size Modelling Using Compound Hawkes Process." *arXiv preprint arXiv:2312.08927*.

Kichler, Neil, Sher Afghan, and Uwe Naumann. 2023. "Towards Sobolev Pruning." *arXiv preprint arXiv:2312.03510*.

Kühn, Christoph, and Christopher Lorenz. 2023. "Insider trading in discrete time Kyle games." *arXiv preprint arXiv:2312.00904*.

Ohnishi, Motoya, Iretiayo Akinola, Jie Xu, Ajay Mandlekar, and Fabio Ramos. 2023. "Signatures Meet Dynamic Programming: Generalizing Bellman Equations for Trajectory Following." *arXiv preprint arXiv:2312.05547*.

Qin, Yiming, Clement Vignac, and Pascal Frossard. 2023. "Sparse Training of Discrete Diffusion Models for Graph Generation." *arXiv preprint arXiv:2311.02142*.

Rahmani, Abdellah, and Pascal Frossard. 2023. "Castor: Causal Temporal Regime Structure Learning." *arXiv preprint arXiv:2311.01412*.

Suder, Piotr M, Jason Xu, and David B Dunson. 2023. "Bayesian Transfer Learning." *arXiv preprint arXiv:2312.13484*.