References with abstracts for QWIM project: machine learning (and more) to forecast time series in quantitative wealth and investment management

Cristian Homescu

December 2022

Abstract

This document includes the list of references (incuding abstracts) for this QWIM project

Contents

1 Motivation for the project

Over recent decades, machine learning ($\mathbb{ML}$) algorithms have achieved remarkable success in various areas. The key to their success is the fact that, given a large representative dataset, $\mathbb{ML}$ algorithms can learn to identify complex non-linear patterns and explore unstructured relationships without hypothesizing them a priori. Thus, $\mathbb{ML}$ algorithms are not limited by assumptions or pre-defined data generating processes, which allows the data to speak for itself.

However, the superiority of $\mathbb{ML}$ is not apparent when it comes to typical time series forecasting, where the data availability is often limited, as shown by results of M4 competition. The strength of $\mathbb{ML}$ algorithms, and in fact the requirement for their successful use, is cross-learning, i.e., using many series to train a single model. This is unlike standard statistical time series algorithms, where a separate model is developed for each series.

Producing probabilistic forecasts for large collections of similar and/or dependent time series is a highly relevant, yet challenging task in practice. While classical time series models fail to capture complex patterns in the data and multivariate techniques struggle to scale to large problem sizes, their reliance on strong structural assumptions makes them data-efficient and allows them to provide estimates of uncertainty. The converse is true for models based on deep neural networks, which can learn complex patterns and dependencies given enough data. A hybrid model that incorporates the benefits of both approaches can deliver better results.

Exploiting cross-series information in forecasting is an idea that gets increased attention lately, especially in the aftermath of the M4 competition. The idea is that instead of developing one model per each time series in the dataset, a global model is developed by exploiting information from many time series simultaneously.

Time series are often hierarchical. Many forecasting applications require multiple time series to be forecast simultaneously. These are often hierarchical in nature and often represent sets of time series which can be highly correlated

Recently, researchers have used the idea of developing global models within a machine learning context. The winning solution at the M4 forecasting competition uses the global model concept with local parameters as well, to cater for individual requirements of different series. This can also be done by clustering groups of related time series. A global model is developed per each cluster.

Performance for time series forecasting has to be assesed through a combination of metrics

It is also argues that it is more practical to compare $\mathbb{ML}$ and statistical forecasting from other perspectives:

- global versus local methods
- probabilistic versus point forecasts
- data driven versus model-driven
- ensemble versus single models
- explanatory/interpretable versus predictive

Combinations of forecasting methods appear to have best results for forecasting of time series of similar granularity as the ones used in $\mathbb{QWIM}$.

1.1 Empirical asset pricing and risk premia forecasting

Asset pricing literature has produced hundreds of potential risk factors. Literature mainly builds on linear or linear-like models, due to simplicity, transparency and computational efficiency. However, this simplicity may miss important features of returns, such as nonlinearity, clustering or dependency structure

The literature shows soncerns for risk premia estimation: potential omission of factors, measurement error, nonlinearity, while linear risk factors may still fall short.

Relative to traditional empirical methods in asset pricing, machine learning accommodates a far more expansive list of potential predictor variables, as well as richer specifications of functional form. Machine learning methods can be successfully applied to the two canonical problems of empirical asset pricing: predicting returns in the cross section and time series.

Gu et al. ("Empirical asset pricing via machine learning," 2020):

- $\mathbb{ML}$ improved performance compared to traditional methods for both cross section and time series stock return prediction

- Predictive gains traced to inclusion of nonlinear interactions
- $\mathbb{ML}$ methods agree on a fairly small set of dominant predictive signals

Leung et al. (“The Promises and Pitfalls of Machine Learning for Predicting Cross-Sectional Stock Returns,” 2020): Despite statistical advantage of $\mathbb{ML}$ model predictions, economic gains corresponding are more limited, and largely dependent on ability to take risk and implement trades efficiently.

The literature has accumulated a staggering list of predictors that various researchers have argued possess forecasting power for returns. The number of stock-level predictive characteristics reported in the literature numbers in the hundreds and macroeconomic predictors of the aggregate market number in the dozens. Additionally, predictors are often close cousins and highly correlated. Traditional prediction methods break down when the predictor count approaches the observation count or predictors are highly correlated. With an emphasis on variable selection and dimension reduction techniques, machine learning is well suited for such challenging prediction problems by reducing degrees of freedom and condensing redundant variation among predictors.

Moreover, there is ambiguity regarding functional forms through which the high-dimensional predictor set enter into risk premia. Should they enter linearly? If nonlinearities are needed, which form should they take? Must we consider interactions among predictors? Such questions rapidly proliferate the set of potential model specifications.

Three aspects of machine learning make it well suited for problems of ambiguous functional form. The first is its diversity. As a suite of dissimilar methods it casts a wide net in its specification search. Second, with methods ranging from generalized linear models to regression trees and neural networks, machine learning is explicitly designed to approximate complex nonlinear associations. Third, parameter penalization and conservative model selection criteria complement the breadth of functional forms spanned by these methods in order to avoid overfit biases and false discovery.

Moreover, macroeconomic signals seem to substantially improve out-of-sample performance, especially when non-linear features are incorporated via machine learning.

2 Relevant references

2.1 Main References

2.1.1 Main references on empirical asset pricing and forecasting of QWIM time series

List of references:

Abhyankar and Wu (“Circus Ring to Zoo to Museum: The Fragility of Factors in Characteristic-based Asset Pricing Models,” 2020)

Baitinger and Flegel (“New Concepts in Financial Forecasting: Network-Based Information, Topological Data Analysis and their Combination,” 2021)

Baltussen et al. (“Predicting Bond Returns: 70 Years of International Evidence,” 2021)

Bessembinder et al. (“Time Series Variation in the Factor Zoo,” 2022)

Bianchi et al. (“Bond Risk Premiums with Machine Learning,” 2021)

Bielinski and Broby (“Machine Learning Methods in Asset Pricing,” 2021)

Breitung and Knuppel (“How far can we forecast? Statistical tests of the predictive content,” 2021)

Bryzgalova et al. (“Bayesian solutions for the factor zoo: we just ran two quadrillion models,” 2021)

Chatigny et al. (“Neural forecasting at scale,” 2021)

Chen and Zimmermann (“Open Source Cross-Sectional Asset Pricing,” 2021)

Chen et al. (“Deep learning in asset pricing,” 2021)

Chiang et al. (“Modeling the cross-section of stock returns using sensible models in a model pool,” 2021)

Chib et al. (“On Comparing Asset Pricing Models,” 2020)

Cohen et al. (“Visual Time Series Forecasting: An Image-driven Approach,” 2021)

Coqueret and Guida (*Machine Learning for Factor Investing: R Version*, 2020)

Czasonis et al. (“Addition by Subtraction: A Better Way to Forecast Factor Returns (and Everything Else),” 2020)

Czasonis et al. (“Relevance,” 2021)

Czasonis et al. (“The Past as Prologue: How to Forecast Presidential Elections,” 2021)

Faria and Verona (“Time-frequency forecast of the equity premium,” 2021)

Feng et al. (“Taming the factor zoo: A test of new factors,” 2020)

Freyberger et al. (“Dissecting Characteristics Nonparametrically,” 2020)

Giovannelli et al. (“Forecasting Stock Returns with Large Dimensional Factor Models,” 2021)

Gospodinov and Robotti (“Common pricing across asset classes: Empirical evidence revisited,” 2021)

Gu et al. (“Autoencoder asset pricing models,” 2021)

Gu et al. (“Empirical asset pricing via machine learning,” 2020)

Hassler and Pohle (“Forecasting under Long Memory,” 2021)

Iworiso and Vrontos (“On the Predictability of the Equity Premium Using Deep Learning Techniques,” 2021)

Kozak et al. (“Shrinking the cross-section,” 2020)

Lettau and Pelger (“Factors That Fit the Time Series and Cross-Section of Stock Returns,” 2020)

Leung et al. (“The Promises and Pitfalls of Machine Learning for Predicting Stock Returns,” 2021)

Meligkotsidou et al. (“Out-of-sample equity premium prediction: a complete subset quantile regression approach,” 2021)

McMillan (“Forecasting sector stock market returns,” 2021)

McMillan (“Forecasting U.S. stock returns,” 2021)

Messmer and Audrino (“The Lasso and the Factor Zoo - Expected Returns in the Cross-Section,” 2020)

Nietert and Otto (“Empirical asset pricing: economic significance and economic model evaluation,” 2020)

Oh and Patton (“Better the Devil You Know: Improved Forecasts from Imperfect Models,” 2021)

Rapach et al. (“Industry return predictability: A machine learning approach,” 2019)

Rapach and Zhou (“Time-series and Cross-sectional Stock Return Forecasting: New Machine Learning Methods,” 2020)

Rapach and Zhou (“Asset Pricing: Time-Series Predictability,” 2022)

Reschenhofer et al. (“Evaluation of current research on stock return predictability,” 2020)

Remlinger et al. (“Expert Aggregation for Financial Forecasting,” 2022)

Ruan et al. (“Stock Price Prediction Under Anomalous Circumstances,” 2021)

Smith and Timmermann (“Break Risk,” 2021)

Smith et al. (“Equity Premium Forecasts with an Unknown Number of Structural Breaks,” 2020)

Wang et al. (“The Best of Both Worlds: Forecasting US Equity Market Returns Using a Hybrid Machine Learning Time Series Approach,” 2021)

Wang et al. (“Forecasting stock returns: A time-dependent weighted least squares approach,” 2021)

Weigand (“Machine learning in empirical asset pricing,” 2019)

Yang et al. (“Why Existing Machine Learning Methods Fails At Extracting the Information of Future Returns Out of Historical Stock Prices : the Curve-Shape-Feature and Non-Curve-Shape-Feature Modes,” 2021)

Yin (“Equity premium prediction: keep it sophisticatedly simple,” 2021)

2.1.2 Main references on time series forecasting

List of references:

Alexander et al. (“Evaluating the Discrimination Ability of Proper Multivariate Scoring Rules,” 2021)

Alexandrov et al. (“GluonTS: Probabilistic and Neural Time Series Modeling in Python,” 2020)

Ansari et al. (“Deep Explicit Duration Switching Models for Time Series,” 2021)

Athanasopoulos and Kourentzes (*On the evaluation of hierarchical forecasts*, 2020)

Atiya (“Why does forecast combination work so well?” 2020)

Bandara et al. (“Improving the accuracy of global forecasting models using time series data augmentation,” 2021)

Bandara et al. (“Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach,” 2020)

Bandara et al. (“Improving the accuracy of global forecasting models using time series data augmentation,” 2021)

Benidis et al. (“Neural forecasting: Introduction and literature overview,” 2020)

Bauer et al. (“Telescope: An Automatic Feature Extraction and Transformation Approach for Time Series Forecasting on a Level-Playing Field,” 2020)

Botchkarev (“A new typology design of performance metrics to measure errors in machine learning regression algorithms,” 2019)

Castle et al. (“Forecasting Principles from Experience with Forecasting Competitions,” 2021)

Cerqueira et al. (“Evaluating time series forecasting models: an empirical study on performance estimation methods,” 2020)

Dama and Sinoquet (“Analysis and modeling to forecast in time series: a systematic review,” 2021)

Faloutsos et al. (“Forecasting Big Time Series: Theory and Practice,” 2019)

Fosten and Gutknecht (“Horizon confidence sets,” 2021)

Gastinger et al. (“A study on Ensemble Learning for Time Series Forecasting and the need for Meta-Learning,” 2021)

Godahewa et al. (“Ensembles of localised models for time series forecasting,” 2021)

Grazzi et al. (“Meta-Forecasting by combining Global Deep Representations with Local Adaptation,” 2021)

Hannadige et al. (“Forecasting a Nonstationary Time Series Using a Mixture of Stationary and Nonstationary Predictors,” 2021)

Hewamalage et al. (“Recurrent Neural Networks for Time Series Forecasting: Current status and future directions,” 2021)

Hunt (“In-sample tests of predictability are superior to pseudo-out-of-sample tests, even when data mining,” 2022)

Hyndman and Athanasopoulos (*Forecasting: Principles and Practice (Third Edition)*, 2020)

Januschowski et al. (“Forecasting with trees,” 2022)

Kang et al. (“Déjà vu: A data-centric forecasting approach through time series cross-similarity,” 2021)

Kang-et-al-2021e

Lichtendahl and Winkler (“Why do some combinations perform better than others?” 2020)

Loning and Kiraly (“Forecasting with sktime: Designing sktime’s New Forecasting API and Applying It to Replicate and Extend the M4 Study,” 2020)

Martin et al. (“Optimal probabilistic forecasts: When do they work?” 2022)

Masini et al. (“Machine Learning Advances for Time Series Forecasting,” 2021)

Montero-Manso and Hyndman (“Principles and Algorithms for Forecasting Groups of Time Series: Locality and Globality,” 2021)

Neto et al. (“Uncovering regimes in out of sample forecast errors,” 2021)

Nybrant (“On Robust Forecast Combinations With Applications to Automated Forecasting,” 2021)

Oreshkin et al. (“N-BEATS: Neural basis expansion analysis for interpretable time series forecasting,” 2020)

Perron and Yamamoto (“Testing for Changes in Forecasting Performance,” 2021)

Petropoulos et al. (“Forecasting: theory and practice,” 2022)

Post et al. (“Robust optimization of forecast combinations,” 2019)

Qian et al. (“Combining forecasts for universally optimal performance,” 2022)

Quaedvlieg (“Multi-Horizon Forecast Comparison,” 2021)

Radchenko et al. (“Too similar to combine? On negative weights in forecast combination,” 2022)

Rossi (“Forecasting in the Presence of Instabilities: How Do We Know Whether Models Predict Well and How to Improve Them,” 2020)

Salinas et al. (“DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks,” 2020)

Siliverstovs and Wochner (“State-Dependent Evaluation of Predictive Ability,” 2021)

Smyl (“A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting,” 2020)

Tadayon and Iwashita (“Comprehensive Analysis of Time Series Forecasting Using Neural Networks,” 2020)

Talagala et al. (“FFORMPP: Feature-based forecast model performance prediction,” 2022)

Thorarinsdottir (“Forecast evaluation,” 2021)

Wu et al. (“AutoCTS: Automated Correlated Time Series Forecasting – Extended Version,” 2021)

2.2 Comprehensive list of references

2.2.1 Forecasting QWIM time series, empirical asset pricing, and predictability of financial returns

List of references:

Ahmed et al. (“Best of the Best: A Comparison of Factor Models,” 2019)

Alhnaity and Abbod (“A new hybrid financial time series prediction model,” 2020)

Arias-Calluari et al. (“Methods for forecasting the effect of exogenous risks on stock markets,” 2021)

Babiak and Barunik (“Deep Learning, Predictability, and Optimal Portfolio Returns,” 2020)

Bahrami et al. (“Are advanced emerging market stock returns predictable? A regime-switching forecast combination approach,” 2019)

Bailey et al. (“Measurement of Factor Strength: Theory and Practice,” 2020)

Baitinger and Flegel (“New Concepts in Financial Forecasting: Network-Based Information, Topological Data Analysis and their Combination,” 2021)

Bali et al. (“Different Strokes: Return Predictability Across Stocks and Bonds with Machine Learning and Big Data,” 2021)

Baltas and Karyampas (“Forecasting the Equity Risk Premium: The Importance of Regime-Dependent Evaluation,” 2020)

Baltussen et al. (“Predicting Bond Returns: 70 Years of International Evidence,” 2020)

Bektic et al. (“Factor-based investing in government bond markets: a survey of the current state of research,” 2020)

Bessembinder et al. (“Time Series Variation in the Factor Zoo,” 2022)

Bianchi and McAlinn (“Divide and Conquer: Financial Ratios and Industry Returns Predictability,” 2021)

Bianchi et al. (“Bond Risk Premiums with Machine Learning,” 2021)

Bianchi and Tamoni (“Sparse Predictive Regressions: Statistical Performance and Economic Significance,” 2020)

Bielinski and Broby (“Machine Learning Methods in Asset Pricing,” 2021)

Blitz et al. (“Five Concerns with the Five-Factor Model,” 2018)

Bryzgalova et al. (“Bayesian solutions for the factor zoo: we just ran two quadrillion models,” 2021)

Capolongo and Pacella (“Forecasting inflation in the euro area: countries matter!” 2021)

Carr and Wu (*Decomposing Long Bond Returns: A Decentralized Theory*, 2021)

Castilho et al. (“Forecasting Financial Market Structure from Network Features using Machine Learning,” 2021)

Charles et al. (“Stock return predictability: Evaluation based on interval forecasts,” 2022)

Chatterjee et al. (“Stock Price Prediction Using Time Series, Econometric, Machine Learning, and Deep Learning Models,” 2021)

Chen and Zimmermann (“Open Source Cross-Sectional Asset Pricing,” 2021)

Cheng et al. (“Financial time series forecasting with multi-modality graph neural network,” 2022)

Chevallier et al. (“Is It Possible to Forecast the Price of Bitcoin?” 2021)

Chiang et al. (“Modeling the cross-section of stock returns using sensible models in a model pool,” 2021)

Chib et al. (“Winners from Winners: A Tale of Risk Factors,” 2022)

Chib et al. (“On Comparing Asset Pricing Models,” 2020)

Chu (“Forecasting Recessions with Financial Variables and Temporal Dependence,” 2021)

Chudik et al. (“Variable Selection and Forecasting in High Dimensional Linear Regressions with Structural Breaks,” 2021)

Cohen et al. (“Visual Forecasting of Time Series with Image-to-Image Regression,” 2020)

Cohen et al. (“Visual Time Series Forecasting: An Image-driven Approach,” 2021)

Collot and Hemauer (“A literature review of new methods in empirical asset pricing: omitted-variable and errors-in-variable bias,” 2021)

Cong et al. (“Deep Sequence Modeling: Development and Applications in Asset Pricing,” 2021)

Cornell (“Stock characteristics and stock returns: a skeptic’s look at the cross section of expected returns,” 2020)

Czasonis et al. (“Addition by Subtraction: A Better Way to Forecast Factor Returns (and Everything Else),” 2020)

Czasonis et al. (“Relevance,” 2021)

Czasonis et al. (“The Past as Prologue: How to Forecast Presidential Elections,” 2021)

Dai et al. (“Predicting stock returns: A risk measurement perspective,” 2021)

Dai et al. (“Forecasting stock return volatility: The role of shrinkage approaches in a data-rich environment,” 2022)

Dendramis et al. (“A similarity-based approach for macroeconomic forecasting,” 2020)

Dong et al. (“Predictive power of ARIMA models in forecasting equity returns: a sliding window method,” 2020)

Dong et al. (“Anomalies and the expected market return,” 2022)

Drobetz and Otto (“Empirical Asset Pricing via Machine Learning: Evidence from the European Stock Market,” 2020)

Elkamhi et al. (“Factor Investing Using Capital Market Assumptions,” 2021)

Ellingsen et al. (“News media vs. FRED-MD for macroeconomic forecasting,” 2022)

Fama and French (“Choosing factors,” 2018)

Fama and French (“Comparing Cross-Section and Time-Series Factor Models,” 2020)

Fan et al. (“FarmTest: Factor-adjusted robust multiple testing with approximate false discovery control,” 2019)

Faria and Verona (“Time-frequency forecast of the equity premium,” 2021)

Fleiss and Cui (“Forecasting Stock Price Changes Using Natural Language Processing,” 2021)

Fulton and Hubrich (“Forecasting US Inflation in Real Time,” 2021)

Gafka et al. (“Sources of Return Predictability,” 2021)

Geertsema and Lu (“Long-horizon predictability and information decay in equity markets,” 2021)

Ghorbani and Chong (“A dimension reduction method for stock-price prediction using multiple predictors,” 2021)

Giovannelli et al. (“Forecasting stock returns with large dimensional factor models,” 2021)

Goliński and Spencer (“Estimating the Term Structure with Linear Regressions: Getting to the Roots of the Problem,” 2021)

Gospodinov and Maasoumi (“Generalized aggregation of misspecified models: with an application to asset pricing,” 2021)

Gospodinov and Robotti (“Common pricing across asset classes: Empirical evidence revisited,” 2021)

Gu et al. (“Autoencoder asset pricing models,” 2021)

Hammerschmid and Lohre (“Regime Shifts and Stock Return Predictability,” 2018)

Haase and Neuenkirch (“Forecasting Stock Market Recessions in the US: Predictive Modeling using Different Identification Approaches,” 2021)

Harvey et al. (“Real-Time Detection of Regimes of Predictability in the U.S. Equity Premium,” 2021)

Hassler and Pohle (“Forecasting under Long Memory,” 2021)

Hauzenberger et al. (“Real-time Inflation Forecasting Using Non-linear Dimension Reduction Techniques,” 2021)

He and Gu (“Multi-modal Attention Network for Stock Movements Prediction,” 2022)

Ho and Lin (“Training by Rolling: Machine Learning and Stock Returns Forecasting,” 2021)

Hull and Qiao (“A Practitioner’s Defense of Return Predictability,” 2017)

Ilmanen et al. (“Demystifying illiquid assets: expected returns for private equity,” 2020)

Iworiso and Vrontos (“On the Predictability of the Equity Premium Using Deep Learning Techniques,” 2021)

Kalfa and Marquez (“Forecasting FOMC Forecasts,” 2021)

Karolyi and Van Nieuwerburgh (“New Methods for the Cross-Section of Returns,” 2020)

Kelly et al. (“Characteristics are covariances: A unified model of risk and return,” 2019)

Klingberg Malmer and Pettersson (“Tidying up the factor zoo: Using machine learning to find sparse factor models that predict asset returns,” 2020)

Kozak et al. (“Shrinking the cross-section,” 2020)

Kynigakis and Panopoulou (“Does Model Complexity add Value to Asset Allocation? Evidence from Machine Learning Forecasting Models,” 2021)

Lee and Seregina (“Optimal Portfolio Using Factor Graphical Lasso,” 2022)

Lettau and Pelger (“Factors That Fit the Time Series and Cross-Section of Stock Returns,” 2020)

Leung et al. (“The Promises and Pitfalls of Machine Learning for Predicting Stock Returns,” 2021)

Li and Bastos (“Stock Market Forecasting Using Deep Learning and Technical Analysis: A Systematic Review,” 2020)

Martinez et al. (*Smooth Robust Multi-Horizon Forecasts*, 2020)

McMillan (“Forecasting sector stock market returns,” 2021)

McMillan (“Forecasting U.S. stock returns,” 2021)

Meligkotsidou et al. (“Out-of-sample equity premium prediction: a complete subset quantile regression approach,” 2021)

Neri (“Domain Specific Concept Drift Detectors for Predicting Financial Time Series,” 2021)

Nevasalmi (“Recession forecasting with high-dimensional data,” 2022)

Nietert and Otto (“Empirical asset pricing: economic significance and economic model evaluation,” 2020)

Noguer i Alonso and Srivastava (“The Shape of Performance Curve in Financial Time Series,” 2021)

Noguer i Alonso et al. (“Deep Learning for Equity Time Series Prediction,” 2020)

Nonejad (“Bayesian model averaging and the conditional volatility process: an application to predicting aggregate equity returns by conditioning on economic variables,” 2021)

Oh and Patton (“Better the Devil You Know: Improved Forecasts from Imperfect Models,” 2021)

Paranhos (“Predicting Inflation with Neural Networks,” 2021)

Pinho (“Forecast comparison of volatility models and their combinations (FTSE100): a tied race,” 2020)

Rahimikia and Poon (“Machine Learning for Realised Volatility Forecasting,” 2021)

Rapach et al. (“Industry return predictability: A machine learning approach,” 2019)

Rapach and Zhou (“Time-series and Cross-sectional Stock Return Forecasting: New Machine Learning Methods,” 2020)

Rapach and Zhou (“Asset Pricing: Time-Series Predictability,” 2022)

Reschenhofer et al. (“Evaluation of current research on stock return predictability,” 2020)

Roy (“A six-factor asset pricing model: The Japanese evidence,” 2021)

Salisu and Tchankam (“US Stock return predictability with high dimensional models,” 2022)

Smith and Timmermann (“Break Risk,” 2021)

Smith et al. (“Equity Premium Forecasts with an Unknown Number of Structural Breaks,” 2020)

Son and Lee (“Graph-based multi-factor asset pricing model,” 2022)

Stein (“Out-of-Sample Equity Premium Prediction: Combination Forecasts with Frequency-Decomposed Variables,” 2021)

Stivers (“Equity premium predictions with many predictors: A risk-based explanation of the size and value factors,” 2018)

Stoyanov and Fabozzi (“Dynamics of Equity Factor Returns and Asset Pricing,” 2021)

Tilly and Livan (“Macroeconomic forecasting with statistically validated knowledge graphs,” 2021)

Tilly et al. (“Macroeconomic forecasting through news, emotions and narrative,” 2021)

Timmermann (“Forecasting methods in finance,” 2018)

Trucíos et al. (“Forecasting Conditional Covariance Matrices in High-Dimensional Time Series: a General Dynamic Factor Approach,” 2021)

Viswanathan and Stephen (“Does Machine Learning Algorithms Improve Forecasting Accuracy? Predicting Stock Market Index Using Ensemble Model,” 2020)

Wang et al. (“The Best of Both Worlds: Forecasting US Equity Market Returns Using a Hybrid Machine Learning Time Series Approach,” 2021)

Wang et al. (“Forecasting stock returns: A time-dependent weighted least squares approach,” 2021)

Weigand (“Machine learning in empirical asset pricing,” 2019)

Wu et al. (“Equity2Vec: End-to-end Deep Learning Framework for Cross-sectional Asset Pricing,” 2021)

Xu et al. (“HIST: A Graph-based Framework for Stock Trend Forecasting via Mining Concept-Oriented Shared Information,” 2022)

Yang et al. (“NumHTML: Numeric-Oriented Hierarchical Transformer Model for Multi-task Financial Forecasting,” 2022)

Yang et al. (“Why Existing Machine Learning Methods Fails At Extracting the Information of Future Returns Out of Historical Stock Prices : the Curve-Shape-Feature and Non-Curve-Shape-Feature Modes,” 2021)

Yara et al. (“Value return predictability across asset classes and commonalities in risk premia,” 2021)

Yin (“Equity premium prediction: keep it sophisticatedly simple,” 2021)

Zeng et al. (“Deep Video Prediction for Time Series Forecasting,” 2021)

Zhan and Xiao (“A Fast Evidential Approach for Stock Forecasting,” 2021)

Zhang (“Empirical asset pricing and ensemble machine learning,” 2021)

Zhu et al. (“High-Dimensional Estimation, Basis Assets, and the Adaptive Multi-Factor Model,” 2021)

2.2.2 Testing procedures for QWIM time series, empirical asset pricing and predictability of financial returns

List of references:

Ardia and Dufays (“Measuring uncertainty and uncertainty dispersion from a large set of model predictions,” 2021)

Barillas and Shanken (“Real-time Portfolio Choice Implications of Asset Pricing Models,” 2019)

Barras (“A large-scale approach for evaluating asset pricing models,” 2019)

Bryzgalova et al. (“Bayesian solutions for the factor zoo: we just ran two quadrillion models,” 2021)

Cai et al. (“Testing capital asset pricing models using functional-coefficient panel data models with cross-sectional dependence,” 2022)

Chai et al. (“Which model best explains the returns of large Australian stocks?” 2019)

Charles et al. (“Stock return predictability: Evaluation based on interval forecasts,” 2022)

Chen et al. (“Predicting returns out of sample: A naive model averaging approach,” 2020)

Chiah et al. (“A Better Model? An Empirical Investigation of the Fama-French Five-factor Model in Australia,” 2016)

Chib and Zeng (“Which factors are risk factors in asset pricing? A model scan framework,” 2020)

Chib et al. (“On Comparing Asset Pricing Models,” 2020)

Chib et al. (“Winners from Winners: A Tale of Risk Factors,” 2022)

Chordia et al. (“Anomalies and false rejections,” 2020)

Frenkel et al. (“Testing for the rationality of central bank interest rate forecasts,” 2021)

Gospodinov and Robotti (“Common pricing across asset classes: Empirical evidence revisited,” 2021)

Goyal and Jegadeesh (“Cross-Sectional and Time-Series Tests of Return Predictability: What Is the Difference?” 2018)

Gramespacher and Banziger (“The Bias in Two-Pass Regression Tests of Asset-Pricing Models in Presence of Idiosyncratic Errors with Cross-Sectional Dependence,” 2019)

Harvey and Liu (“Detecting Repeatable Performance,” 2020)

Harvey et al. (“An Evaluation of Alternative Multiple Testing Methods for Finance Applications,” 2020)

Hoga and Dimitriadis (“On Testing Equal Conditional Predictive Ability Under Measurement Error,” 2021)

Janssen (“Multi-horizon comparison of multivariate inflation forecasting,” 2019)

Jegadeesh et al. (“Empirical tests of asset pricing models with individual assets: Resolving the errors-in-variables bias in risk premium estimation,” 2019)

Kelly et al. (“Characteristics are covariances: A unified model of risk and return,” 2019)

Kruse et al. (“Comparing Predictive Accuracy under Long Memory, With an Application to Volatility Forecasting,” 2019)

Kyriakou et al. (“Longer-Term Forecasting of Excess Stock Returns – The Five-Year Case,” 2020)

Ledoit et al. (“Efficient Sorting: A More Powerful Test for Cross-Sectional Anomalies,” 2019)

Mikeliani and Kavlashvili (“Evaluation and comparison of machine learning and classical econometric AR model on financial time series data,” 2020)

Odendahl et al. (“Comparing Forecast Performance with State Dependence,” 2020)

Pesaran and Smith (“The Role of Factor Strength and Pricing Errors for Estimation and Inference in Asset Pricing Models,” 2019)

Pinho (“Forecast comparison of volatility models and their combinations (FTSE100): a tied race,” 2020)

Prasad et al. (“Prediction of Stock Prices Using Statistical and Machine Learning Models: A Comparative Analysis,” 2021)

Siami-Namini et al. (“A Comparative Analysis of Forecasting Financial Time Series Using ARIMA, LSTM, and BiLSTM,” 2019)

Siliverstovs and Wochner (“State-Dependent Evaluation of Predictive Ability,” 2021)

Suhonen et al. (“Quantifying Backtest Overfitting in Alternative Beta Strategies,” 2017)

Tang et al. (“Out-of-sample equity premium prediction: A scenario analysis approach,” 2018)

Tunaru et al. (“Testing the Forecasting Ability of Multi-Factor Models on Non-US Interbank Rates,” 2021)

Vincent et al. (“Investment styles and the multiple testing of cross-sectional stock return predictability,” 2020)

Xie (“Forecasting Long-Term Equity Returns: A Comparison of Popular Methodologies,” 2021)

Zhao (“Essays on Asset Pricing: A Model Comparison Perspective,” 2020)

2.2.3 Forecasting time series

List of references:

Alexandrov et al. (“GluonTS: Probabilistic and Neural Time Series Modeling in Python,” 2020)

Ankile and Krange (“The DONUT Approach to Ensemble Combination Forecasting,” 2022)

Ansari et al. (“Deep Explicit Duration Switching Models for Time Series,” 2021)

Ashouri et al. (“Fast Forecast Reconciliation Using Linear Models,” 2022)

Athanasopoulos and Kourentzes (*On the evaluation of hierarchical forecasts*, 2020)

Athanasopoulos et al. (“Hierarchical Forecasting,” 2019)

Bandara et al. (“Improving the accuracy of global forecasting models using time series data augmentation,” 2021)

Berardi et al. (“Mind the (Convergence) Gap: Bond Predictability Strikes Back!” 2021)

Berrisch and Ziel (“CRPS Learning,” 2021)

Bisaglia and Grigoletto (“A new time-varying model for forecasting long-memory series,” 2021)

Castle et al. (“Forecasting Principles from Experience with Forecasting Competitions,” 2021)

Castle et al. (“Selecting a Model for Forecasting,” 2021)

Cerqueira et al. (“Evaluating time series forecasting models: an empirical study on performance estimation methods,” 2020)

Cerqueira et al. (“Model Selection for Time Series Forecasting: Empirical Analysis of Different Estimators,” 2021)

Challu et al. (“DMIDAS: Deep Mixed Data Sampling Regression for Long Multi-Horizon Time Series Forecasting,” 2021)

Chen et al. (“Learning from Multiple Time Series: A Deep Disentangled Approach to Diversified Time Series Forecasting,” 2021)

Dama and Sinoquet (“Analysis and modeling to forecast in time series: a systematic review,” 2021)

Deshpande et al. (“Long Horizon Forecasting With Temporal Point Processes,” 2021)

Faloutsos et al. (“Forecasting Big Time Series: Theory and Practice,” 2019)

Feldkircher et al. (“Factor Augmented Vector Autoregressions, Panel VARs, and Global VARs,” 2020)

Godahewa et al. (“Monash Time Series Forecasting Archive,” 2021)

Grazzi et al. (“Meta-Forecasting by combining Global Deep Representations with Local Adaptation,” 2021)

Harris et al. (“Construction and visualization of confidence sets for frequentist distributional forecasts,” 2019)

Hassler and Pohle (“Forecasting under Long Memory,” 2021)

Hewamalage et al. (“Global Models for Time Series Forecasting: A Simulation Study,” 2020)

Hewamalage et al. (“Recurrent Neural Networks for Time Series Forecasting: Current status and future directions,” 2021)

Hyndman and Athanasopoulos (*Forecasting: Principles and Practice (Third Edition)*, 2020)

Inoue et al. (“Rolling window selection for out-of-sample forecasting with time-varying parameters,” 2017)

Kang et al. (“Déjà vu: A data-centric forecasting approach through time series cross-similarity,” 2021)

Liu et al. (“Forecast Methods for Time Series Data: A Survey,” 2021)

Loning and Kiraly (“Forecasting with sktime: Designing sktime’s New Forecasting API and Applying It to Replicate and Extend the M4 Study,” 2020)

Makridakis et al. (“Forecasting in social settings: The state of the art,” 2020)

Masini et al. (“Machine Learning Advances for Time Series Forecasting,” 2021)

Montero-Manso and Hyndman (“Principles and Algorithms for Forecasting Groups of Time Series: Locality and Globality,” 2020)

Nystrup et al. (“Dimensionality reduction in forecasting with temporal hierarchies,” 2021)

Petropoulos and Grushka-Cockayne (“Fast and Frugal Time Series Forecasting,” 2021)

Petropoulos et al. (“Forecasting: theory and practice,” 2022)

Okuno et al. (“Combining multiple forecasts for multivariate time series via state-dependent weighting.,” 2019)

Oreshkin et al. (“N-BEATS: Neural basis expansion analysis for interpretable time series forecasting,” 2019)

Salinas et al. (“DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks,” 2020)

Smyl (“A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting,” 2020)

Tadayon and Iwashita (“Comprehensive Analysis of Time Series Forecasting Using Neural Networks,” 2020)

Talagala et al. (“FFORMPP: Feature-based forecast model performance prediction,” 2022)

Timmermann (“Forecasting methods in finance,” 2018)

Wang et al. (“Deep Factors for Forecasting,” 2019)

Wellens et al. (“Transfer learning for hierarchical forecasting: Reducing computational efforts of M5 winning methods,” 2022)

Wen et al. (“Forecasting realized volatility of Chinese stock market: A simple but efficient truncated approach,” 2022)

Wu et al. (“AutoCTS: Automated Correlated Time Series Forecasting – Extended Version,” 2021)

2.2.4 Forecasting time series using Machine Learning

List of references:

Alexandrov et al. (“GluonTS: Probabilistic and Neural Time Series Modeling in Python,” 2020)

Ansari et al. (“Deep Explicit Duration Switching Models for Time Series,” 2021)

Babii et al. (“Machine Learning Panel Data Regressions with Heavy-tailed Dependent Data: Theory and Application,” 2021)

Bhatnagar et al. (“Merlion: A Machine Learning Library for Time Series,” 2021)

Bielinski and Broby (“Machine Learning Methods in Asset Pricing,” 2021)

Bloemheuvel et al. (“Multivariate Time Series Regression with Graph Neural Networks,” 2022)

Castilho et al. (“Forecasting Financial Market Structure from Network Features using Machine Learning,” 2021)

Chen et al. (“Deep learning in asset pricing,” 2021)

Chen et al. (“Multi-Scale Adaptive Graph Neural Network for Multivariate Time Series Forecasting,” 2022)

Chatterjee et al. (“Stock Price Prediction Using Time Series, Econometric, Machine Learning, and Deep Learning Models,” 2021)

Chatigny et al. (“Neural forecasting at scale,” 2021)

Cholakov and Kolev (“Transformers predicting the future. Applying attention in next-frame and time series forecasting,” 2021)

Cohen et al. (“Visual Time Series Forecasting: An Image-driven Approach,” 2021)

Debnath et al. (“Exploring Generative Data Augmentation in Multivariate Time Series Forecasting : Opportunities and Challenges,” 2021)

Du et al. (“AdaRNN: Adaptive Learning and Forecasting of Time Series,” 2021)

Faloutsos et al. ("Forecasting Big Time Series: Theory and Practice," 2019)

Filipovic and Khalilzadeh ("Machine Learning for Predicting Stock Return Volatility," 2021)

Fjellstrom ("Long Short-Term Memory Neural Network for Financial Time Series," 2022)

Geertsema and Lu ("Long-horizon predictability and information decay in equity markets," 2021)

Harris et al. ("Construction and visualization of confidence sets for frequentist distributional forecasts," 2019)

Herzen et al. ("Darts: User-Friendly Modern Machine Learning for Time Series," 2022)

Hewamalage et al. ("Recurrent Neural Networks for Time Series Forecasting: Current status and future directions," 2021)

Januschowski et al. ("Forecasting with trees," 2022)

Jin et al. ("Robust Forecast Comparison," 2017)

Kiefer et al. ("Univariate Time Series Forecasting: Machine Learning Prediction of the Best Suitable Forecast Model Based on Time Series Characteristics," 2021)

Kosman and Castro ("Vision-Guided Forecasting – Visual Context for Multi-Horizon Time Series Forecasting," 2021)

Kynigakis and Panopoulou ("Does Model Complexity add Value to Asset Allocation? Evidence from Machine Learning Forecasting Models," 2021)

Lara-Benítez et al. ("Evaluation of the Transformer Architecture for Univariate Time Series Forecasting," 2021)

Lara-Benitez et al. ("An Experimental Review on Deep Learning Architectures for Time Series Forecasting," 2021)

Le Guen and Thome ("Deep Time Series Forecasting with Shape and Temporal Criteria," 2021)

Li and Bastos ("Stock Market Forecasting Using Deep Learning and Technical Analysis: A Systematic Review," 2020)

Li et al. ("Bayesian forecast combination using time-varying features," 2021)

Lim and Zohren ("Time-series forecasting with deep learning: a survey," 2021)

Liu et al. ("Time Series is a Special Sequence: Forecasting with Sample Convolution and Interaction," 2021)

Loning and Kiraly ("Forecasting with sktime: Designing sktime's New Forecasting API and Applying It to Replicate and Extend the M4 Study," 2020)

Mancuso et al. ("A machine learning approach for forecasting hierarchical time series," 2021)

Masini et al. ("Machine Learning Advances for Time Series Forecasting," 2021)

Nevasalmi ("Forecasting multinomial stock returns using machine learning methods," 2020)

Noguer i Alonso and Srivastava ("The Shape of Performance Curve in Financial Time Series," 2021)

Papaioannou et al. ("Time Series Forecasting Using Manifold Learning," 2021)

Paranhos ("Predicting Inflation with Neural Networks," 2021)

Oreshkin et al. ("N-BEATS: Neural basis expansion analysis for interpretable time series forecasting," 2019)

Petropoulos and Spiliotis ("The Wisdom of the Data: Getting the Most Out of Univariate Time Series Forecasting," 2021)

Petropoulos et al. ("Forecasting: theory and practice," 2022)

Pinto and Castle (*A machine learning dynamic switching approach to forecasting when there are structural breaks*, 2021)

Rajapaksha et al. ("LoMEF: A Framework to Produce Local Explanations for Global Model Time Series Forecasts," 2021)

Rožanec et al. ("Explaining Bad Forecasts in Global Time Series Models," 2021)

Salinas et al. ("DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks," 2020)

Smyl ("A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting," 2020)

Spiliotis et al. ("Hierarchical forecast reconciliation with machine learning," 2021)

Tadayon and Iwashita ("Comprehensive Analysis of Time Series Forecasting Using Neural Networks," 2020)

Theodosiou and Kourentzes ("Forecasting with Deep Temporal Hierarchies," 2021)

Viswanathan and Stephen ("Does Machine Learning Algorithms Improve Forecasting Accuracy? Predicting Stock Market Index Using Ensemble Model," 2020)

Wang et al. ("Deep Factors for Forecasting," 2019)

Wellens et al. ("Transfer learning for hierarchical forecasting: Reducing computational efforts of M5 winning methods," 2022)

Wen et al. (“Forecasting realized volatility of Chinese stock market: A simple but efficient truncated approach,” 2022)

Wu et al. (“Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting,” 2022)

Xu et al. (“Instance-wise Graph-based Framework for Multivariate Time Series Forecasting,” 2021)

Xu et al. (“HIST: A Graph-based Framework for Stock Trend Forecasting via Mining Concept-Oriented Shared Information,” 2022)

Yang et al. (“Why Existing Machine Learning Methods Fails At Extracting the Information of Future Returns Out of Historical Stock Prices : the Curve-Shape-Feature and Non-Curve-Shape-Feature Modes,” 2021)

Zhang (“Empirical asset pricing and ensemble machine learning,” 2021)

2.2.5 Testing procedures to evaluate and compare forecasts

List of references:

Anghel (“Data Snooping Bias in Tests of the Relative Performance of Multiple Forecasting Models,” 2021)

Aparicio and Lopez de Prado (“How Hard Is It to Pick the Right Model? MCS and backtest overfitting,” 2018)

Arnold et al. (“Sequentially valid tests for forecast calibration,” 2022)

Arvanitis et al. (“Nonparametric tests for Optimal Predictive Ability,” 2021)

Bates et al. (“Cross-validation: what does it estimate and how well does it do it?” 2021)

Ben Baccar (“Comparative Study on Time Series Forecasting Models,” 2019)

Bergmeir et al. (“A note on the validity of cross-validation for evaluating autoregressive time series prediction,” 2018)

Bouallegue et al. (“The diagonal score: Definition, properties, and interpretations,” 2018)

Brehmer and Gneiting (“Properization: constructing proper scoring rules via Bayes acts,” 2020)

Brehmer et al. (“Using scoring functions to evaluate point process forecasts,” 2021)

Breitung and Knuppel (“How far can we forecast? Statistical tests of the predictive content,” 2021)

Bulut (“Does Statistical Significance Help to Evaluate Predictive Performance of Competing Models?” 2019)

Cerqueira et al. (“Model Selection for Time Series Forecasting: Empirical Analysis of Different Estimators,” 2021)

Cetin and Yavuz (“Comparison of forecast accuracy of Ata and exponential smoothing,” 2021)

Choe and Ramdas (“Comparing Sequential Forecasters,” 2022)

Coroneo et al. (“Testing the Predictive Accuracy of COVID-19 Forecasts,” 2021)

Costantini and Kunst (“On using predictive-ability tests in the selection of time-series prediction models: A Monte Carlo evaluation,” 2021)

Davydenko and Goodwin (“Assessing Point Forecast Bias Across Multiple Time Series: Measures and Visual Tools,” 2021)

De Baets and Harvey (“Using judgment to select and adjust forecasts from statistical models,” 2020)

Diebold (“Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold-Mariano Tests,” 2015)

Fauvel et al. (“A Performance-Explainability Framework to Benchmark Machine Learning Methods: Application to Multivariate Time Series Classifiers,” 2021)

Fosten and Gutknecht (“Horizon confidence sets,” 2021)

Geweke and Amisano (“Comparing and evaluating Bayesian predictive distributions of asset returns,” 2010)

Gilleland et al. (“Testing the Tests: What Are the Impacts of Incorrect Assumptions When Applying Confidence Intervals or Hypothesis Tests to Compare Competing Forecasts?” 2018)

Gneiting and Resin (“Regression Diagnostics meets Forecast Evaluation: Conditional Calibration, Reliability Diagrams, and Coefficient of Determination,” 2022)

Hounyo and Lahiri (“Estimating the variance of a combined forecast: Bootstrap-based approach,” 2022)

Hunt (“In-sample tests of predictability are superior to pseudo-out-of-sample tests, even when data mining,” 2022)

Ilic et al. (“Augmented Out-of-Sample Comparison Method for Time Series Forecasting Techniques,” 2020)

Jin et al. (“Robust Forecast Comparison,” 2017)

Kang et al. (“Assessing Goodness of Fit for Verifying Probabilistic Forecasts,” 2021)

Koutsandreas et al. (“On the selection of forecasting accuracy measures,” 2021)

Lerch et al. ("Forecaster's Dilemma: Extreme Events and Forecast Evaluation," 2015)
Lerch et al. ("Forecaster's Dilemma: Extreme Events and Forecast Evaluation," 2017)
Martin et al. ("Optimal probabilistic forecasts: When do they work?" 2022)
McCracken ("Tests of Conditional Predictive Ability: Existence, Size, and Power," 2020)
Murray and Blume ("False Discovery Rate Computation: Illustrations and Modifications," 2020)
Neto et al. ("Uncovering regimes in out of sample forecast errors," 2021)
Patton ("Comparing Possibly Misspecified Forecasts," 2020)
Perron and Yamamoto ("Testing for Changes in Forecasting Performance," 2021)
Pitarakis ("A Novel Approach to Predictive Accuracy Testing in Nested Environments," 2020)
Qu et al. ("Comparing forecasting performance in cross-sections," 2021)
Quaedvlieg ("Multi-Horizon Forecast Comparison," 2021)
Rožanec et al. ("Explaining Bad Forecasts in Global Time Series Models," 2021)
Rytchkov and Zhong ("Information Aggregation and P-Hacking," 2020)
Sharma et al. ("Prediction-Oriented Model Selection in Partial Least Squares Path Modeling," 2020)
Siliverstovs and Wochner ("State-Dependent Evaluation of Predictive Ability," 2021)
Spiliotis et al. ("Tales from tails: On the empirical distributions of forecasting errors and their implication to risk," 2019)
Stauskas and Westerlund ("Tests of Equal Forecasting Accuracy for Nested Models with Estimated CCE Factors," 2022)
Taggart ("Evaluation of point forecasts for extreme events using consistent scoring functions," 2021)
Taillardat et al. ("Extreme events evaluation using CRPS distributions," 2022)
Vovk and Wang ("E-values: Calibration, combination, and applications," 2021)
Wang and Ramdas ("False discovery rate control with e-values," 2020)
Westerlund et al. ("Testing for Predictability in panels with General Predictors," 2017)
Yeoleka et al. ("Feature Selection on a Flare Forecasting Testbed: A Comparative Study of 24 Methods," 2021)
Zhao et al. ("Empirical Quantitative Analysis of COVID-19 Forecasting Models," 2021)
Zhu and Timmermann ("Can Two Forecasts Have the Same Conditional Expected Accuracy?" 2020)
Ziel and Berk ("Multivariate Forecasting Evaluation: On Sensitive and Strictly Proper Scoring Rules," 2019)

2.2.6 Combinations of forecasting methods for time series

List of references:
Atiya ("Why does forecast combination work so well?" 2020)
Bahrami et al. ("Are advanced emerging market stock returns predictable? A regime-switching forecast combination approach," 2019)
Caldeira et al. ("Yield curve forecast combinations based on bond portfolio performance," 2018)
Cerqueira et al. ("Model Compression for Dynamic Forecast Combination," 2021)
Chan and Pauwels ("Some theoretical results on forecast combinations," 2018)
Di Fonzo and Girolimetto ("Cross-temporal forecast reconciliation: Optimal combination method and heuristic alternatives," 2020)
Di Fonzo and Girolimetto ("Forecast combination based forecast reconciliation: insights and extensions," 2021)
Di Fonzo and Girolimetto ("Cross-temporal forecast reconciliation: Optimal combination method and heuristic alternatives," 2022)
Fameliti and Skintzi ("Predictive ability and economic gains from volatility forecast combinations," 2020)
Fang et al. ("Optimal forecast combination based on ensemble empirical mode decomposition for agricultural commodity futures prices," 2020)
Godahewa et al. ("Ensembles of localised models for time series forecasting," 2021)
Hannadige et al. ("Forecasting a Nonstationary Time Series Using a Mixture of Stationary and Nonstationary Predictors," 2021)
Hofmarcher and Grun ("Bayesian Model Averaging," 2020)
Hollyman et al. ("Understanding forecast reconciliation," 2021)
Holzmann and Klar ("Using Proxies to Improve Forecast Evaluation," 2021)
Hsiao and Wan ("Is there an optimal forecast combination?" 2014)
Jaganathan and Prakash ("A combination-based forecasting method for the M4-competition," 2020)

Lee and Seregina (“Optimal Portfolio Using Factor Graphical Lasso,” 2022)

Lichtendahl and Winkler (“Why do some combinations perform better than others?” 2020)

Montero-Manso et al. (“FFORMA: Feature-based Forecast Model Averaging,” 2020)

Montero-Manso and Hyndman (“Principles and Algorithms for Forecasting Groups of Time Series: Locality and Globality,” 2021)

Nybrant (“On Robust Forecast Combinations With Applications to Automated Forecasting,” 2021)

Okuno et al. (“Combining multiple forecasts for multivariate time series via state-dependent weighting.,” 2019)

Patton (“Comparing Possibly Misspecified Forecasts,” 2020)

Petropoulos and Svetunkov (“A simple combination of univariate models,” 2020)

Petropoulos et al. (“Model combinations through revised base-rates,” 2021)

Post et al. (“Robust optimization of forecast combinations,” 2019)

Qian et al. (“On the forecast combination puzzle,” 2019)

Qian et al. (“Combining forecasts for universally optimal performance,” 2022)

Radchenko et al. (“Too similar to combine? On negative weights in forecast combination,” 2022)

Rehman et al. (“Individual and combination approaches to forecasting hierarchical time series with correlated data: an empirical study,” 2019)

Roccazzella et al. (“Optimal and robust combination of forecasts via constrained optimization and shrinkage,” 2022)

Shaub (“Fast and accurate yearly time series forecasting with forecast combinations,” 2020)

Stein (“Out-of-Sample Equity Premium Prediction: Combination Forecasts with Frequency-Decomposed Variables,” 2021)

Svensson (“An Evaluation of Methods for Combining Univariate Time Series Forecasts,” 2018)

Thomson et al. (“Combining forecasts: Performance and coherence,” 2019)

Vaiciukynas et al. (“Two-Step Meta-Learning for Time-Series Forecasting Ensemble,” 2022)

van Dijk and Franses (“Combining expert-adjusted forecasts,” 2019)

Weiss et al. (“Forecast Combinations in R using the ForecastComb Package,” 2018)

Winkler (“Equal Versus Differential Weighting in Combining Forecasts,” 2015)

Zhang (“Empirical asset pricing and ensemble machine learning,” 2021)

Zhao (“The robustness of forecast combination in unstable environments: a Monte Carlo study of advanced algorithms,” 2021)

2.2.7 Combination of statistical and machine learning approaches

List of references:

Allende and Valle (“Ensemble methods for time series forecasting,” 2017)

Barrow and Crone (“A comparison of AdaBoost algorithms for time series forecast combination,” 2016)

Bergmeir et al. (“Time Series Modeling and Forecasting Using Memetic Algorithms for Regime-Switching Models,” 2012)

Billio et al. (“Time-varying combinations of predictive densities using nonlinear filtering,” 2013)

Gilliland (“The value added by machine learning approaches in forecasting,” 2020)

Grazzi et al. (“Meta-Forecasting by combining Global Deep Representations with Local Adaptation,” 2021)

Habibnia (“Essays in high-dimensional nonlinear time series analysis,” 2016)

Hewamalage et al. (“Recurrent Neural Networks for Time Series Forecasting: Current status and future directions,” 2021)

Joshi (“Time Series Analysis and Forecasting of the US Housing Starts using Econometric and Machine Learning Model,” 2019)

Karathanasopoulos et al. (“Modelling and Trading the English and German Stock Markets with Novelty Optimization Techniques,” 2017)

Kuznetsov and Mohri (“Time series prediction and online learning,” 2016)

McDonald et al. (“A comparison of forecasting approaches for capital markets,” 2014)

Menezes and Mastelini (“MegazordNet: combining statistical and machine learning standpoints for time series forecasting,” 2021)

Pinto and Marçal (“Cross-Validation Based Forecasting Method: A Machine Learning Approach,” 2019)

Pinto and Marçal ("Inflation Rate Forecasting: Extreme Learning Machine as a Model Combination Method," 2020)

Qian et al. ("Combining forecasts for universally optimal performance," 2022)

Risse ("Combining Wavelet Decomposition with Machine Learning to Forecast Gold Returns," 2017)

Talagala et al. ("FFORMPP: Feature-based forecast model performance prediction," 2021)

Vaiciukynas et al. ("Two-Step Meta-Learning for Time-Series Forecasting Ensemble," 2022)

Viswanathan and Stephen ("Does Machine Learning Algorithms Improve Forecasting Accuracy? Predicting Stock Market Index Using Ensemble Model," 2020)

Zang ("Deep Learning in Multiple Multistep Time Series Prediction," 2017)

2.2.8 Probabilistic forecasting of time series

List of references:

Bjerregård et al. ("An introduction to multivariate probabilistic forecast evaluation," 2021)

Bouallegue et al. ("The diagonal score: Definition, properties, and interpretations," 2018)

Deshpande and Sarawagi ("Long Range Probabilistic Forecasting in Time-Series using High Order Statistics," 2021)

Gonzalez-Rivera et al. ("Prediction regions for interval-valued time series," 2020)

Graziani et al. ("Probabilistic recalibration of forecasts," 2021)

Greenberg ("Calibration Scoring Rules for Practical Prediction Training," 2020)

Jordan et al. ("Evaluating probabilistic forecasts with scoringRules," 2019)

Le Guen and Thome ("Probabilistic Time Series Forecasting with Structured Shape and Temporal Diversity," 2020)

Lerch et al. ("Forecaster's Dilemma: Extreme Events and Forecast Evaluation," 2017)

Lerch et al. ("Forecaster's Dilemma: Extreme Events and Forecast Evaluation," 2015)

Kamarthi et al. ("CAMul: Calibrated and Accurate Multi-view Time-Series Forecasting," 2021)

Kang et al. ("Assessing Goodness of Fit for Verifying Probabilistic Forecasts," 2021)

Martin et al. ("Optimal probabilistic forecasts: When do they work?" 2020)

Prayogo et al. ("Time Series Sampling for Probabilistic Forecasting," 2020)

Taylor and Taylor ("Combining probabilistic forecasts of COVID-19 mortality in the United States," 2021)

2.2.9 Metrics to assess forecast performance

List of references:

Alexander et al. ("Evaluating the Discrimination Ability of Proper Multivariate Scoring Rules," 2021)

Botchkarev ("A new typology design of performance metrics to measure errors in machine learning regression algorithms," 2019)

Cheng et al. ("Forecast Evaluation," 2019)

Chiu et al. ("A new approach for detecting shifts in forecast accuracy," 2019)

Gasthaus et al. ("Probabilistic Forecasting with Spline Quantile Function RNNs," 2019)

Hyndman and Athanasopoulos (*Forecasting: Principles and Practice (Third Edition)*, 2020)

Makridakis et al. ("The M4 Competition: 100,000 time series and 61 forecasting methods," 2019)

Makridakis and Petropoulos ("The M4 competition: Conclusions," 2020)

Neto et al. ("Uncovering regimes in out of sample forecast errors," 2021)

Petropoulos et al. ("The inventory performance of forecasting methods: Evidence from the M3 competition data," 2019)

Ryll and Seidens ("Evaluating the Performance of Machine Learning Algorithms in Financial Market Forecasting: A Comprehensive Survey," 2019)

Samuels and Sekkel ("Model Confidence Sets and forecast combination," 2017)

Thorarinsdottir ("Forecast evaluation," 2021)

2.2.10 Software implementations and frameworks

List of references:

Alexandrov et al. ("GluonTS: Probabilistic and Neural Time Series Modeling in Python," 2020)

Bacher et al. ("onlineforecast: An R package for adaptive and recursive forecasting," 2022)

Bernardi and Catania ("The Model Confidence Set package for R," 2014)

Bhatnagar et al. ("Merlion: A Machine Learning Library for Time Series," 2021)

Bokde et al. ("ForecastTB - An R Package as a Test-Bench for Time Series Forecasting, with Application of Wind Speed and Solar Radiation Modeling," 2020)

Burns and Whyne ("Seglearn: A Python Package for Learning Sequences and Time Series," 2018)

Charte et al. ("predtoolsTS: R package for streamlining time series forecasting," 2019)

de Valk et al. ("Nowcasting: An R Package for Predicting Economic Variables Using Dynamic Factor Models," 2019)

Golyandina et al. (*Singular Spectrum Analysis with R*, 2018)

Herzen et al. ("Darts: User-Friendly Modern Machine Learning for Time Series," 2022)

Hyndman ("Tidy Time Series and Forecasting in R," 2020)

Jordan et al. ("Evaluating probabilistic forecasts with scoringRules," 2019)

Leroy et al. ("MAGMA: Inference and Prediction with Multi-Task Gaussian Processes," 2020)

Loning et al. ("sktime: A Unified Interface for Machine Learning with Time Series," 2019)

Qian et al. ("Combining forecasts for universally optimal performance," 2022)

Salles et al. ("TSPred: A framework for nonstationary time series prediction," 2022)

Seca ("TimeGym: Debugging for Time Series Modeling in Python," 2021)

Shaub ("Fast and accurate yearly time series forecasting with forecast combinations," 2020)

Siebert et al. ("A systematic review of Python packages for time series analysis," 2021)

References

Abhyankar, A. and Wu, Y. (2020). "Circus Ring to Zoo to Museum: The Fragility of Factors in Characteristic-based Asset Pricing Models." In: *SSRN e-Print*.

Economically relevant factors in asset pricing models should impound information on the future path of state variables that drive asset risk premia. Imposing this condition, we investigate which publicly available characteristics predict individual stock returns during the sample period used by Fama and French (1993) i.e. their discovery period and the post-1993 or out-of-sample period. We find four characteristics have significant predictive power, in the cross-section, over and above that of their factors. In the out-of-sample period, five new characteristics become significant predictors. Similar results are seen for the Chen and Zhang (2010) model. Next, we find that characteristics that forecast stock returns before and after major economic events are very different. Finally, we find that the ability of characteristics to reflect economic uncertainty and sentiment changes in sign and magnitude over time- often vanishing altogether. Our results suggest that the search for a unique characteristic-based asset pricing model is unlikely to be fruitful given the secular variation in the relation between the sources of macroeconomic risks and firm-level characteristics.

Ahmed, S., Bu, Z., and Tsvetanov, D. (2019). "Best of the Best: A Comparison of Factor Models." In: *Journal of Financial and Quantitative Analysis* 54(4), pp. 1713–1758.

We compare major factor models and find that the Stambaugh and Yuan (2016) four-factor model is the overall winner in the time-series domain. The Hou, Xue, and Zhang (2015) q-factor model takes second place and the Fama and French (2015) five-factor model and the Barillas and Shanken (2018) six-factor model jointly take third place. But the pairwise cross-sectional R2 and the multiple model comparison tests show that the Hou, Xue, and Zhang (2015) q-factor model, the Fama and French (2015) five-factor and four-factor models, and the Barillas and Shanken (2018) six-factor model take equal first place in the horse race.

Alexander, C., Coulon, M., Han, Y., and Meng, X. (2021). "Evaluating the Discrimination Ability of Proper Multivariate Scoring Rules." In: *arXiv e-Print*.

Proper scoring rules are commonly applied to quantify the accuracy of distribution forecasts. Given an observation they assign a scalar score to each distribution forecast, with the the lowest expected score attributed to the true distribution. The energy and variogram scores are two rules that have recently gained some popularity in multivariate settings because their computation does not require a forecast to have parametric density function and so they are broadly applicable. Here we conduct a simulation study to compare the discrimination ability between the energy score and three variogram scores. Compared with other studies, our simulation design is more realistic because it is supported by a historical data set containing commodity prices, currencies and interest rates, and our data generating processes include a diverse selection of models with different marginal distributions, dependence structure, and calibration windows. This facilitates a comprehensive comparison of the performance of proper scoring rules in different settings. To compare the scores we use three metrics: the mean relative score, error rate and a generalised discrimination heuristic. Overall, we find that the variogram score with parameter p=0.5 outperforms the energy score and the other two variogram scores.

Alexandrov, A., Benidis, K., Bohlke-Schneider, M., Flunkert, V., Gasthaus, J., Januschowski, T., Maddix, D. C., Rangapuram, S., Salinas, D., Schulz, J., Stella, L., Turkmen, A. C., and Wang, Y. (2020). "GluonTS: Probabilistic and Neural Time Series Modeling in Python." In: *Journal of Machine Learning Research* 21(116), pp. 1–6.

We introduce the Gluon Time Series Toolkit (GluonTS), a Python library for deep learning based time series modeling for ubiquitous tasks, such as forecasting and anomaly detection. GluonTS simplifies the time series modeling pipeline by providing the necessary components and tools for quick model development, efficient experimentation and evaluation. In addition, it contains reference implementations of state-of-the-art time series models that enable simple benchmarking of new algorithms.

Alhnaity, B. and Abbod, M. (2020). "A new hybrid financial time series prediction model." In: *Engineering Applications of Artificial Intelligence* 95, p. 103873.

Due to the characteristics of financial time series, such as being non-linear, non-stationary and noisy, with uncertain and hidden relationships, it is difficult to capture its non-stationary state and to accurately describe its moving tendency. This is also a consequence of using a single approach to artificial intelligence, and other techniques that have been conventionally used. Therefore, those participating in financial markets, along with researchers, have paid a great deal of attention to tackling this problem. Hence, several approaches have been developed to alleviate the influence of inherent characteristics. However, the noise characteristic can refer to the unavailability of information, which affects how financial markets behave, as well as captured prices in both the past and the future. Therefore, the prediction of stock prices and detecting their noise is considered a very challenging financial topic. This paper adopts a novel three-step hybrid intelligent prediction model that combines a collection of intelligent modelling techniques and a feature extraction algorithm. At first, ensemble empirical mode decomposition is applied to the original data, as to facilitate model fitting to them. Then neural network and support vector regression is proposed individually for modelling the extracted features. Finally, a weighted ensemble average using a genetic algorithm to optimise and determine the weight is proposed for establishing a unified prediction. Experimental results are presented which illustrate the excellent performance of the proposed approach, and that is significantly outperforming the existing models, in terms of error criteria such as MSE, RMSE and MAE.

Allende, H. and Valle, C. (2017). "Ensemble methods for time series forecasting." In: *Claudio Moraga: A Passion for Multi-Valued Logic and Soft Computing*. Ed. by R. Seising and H. Allende-Cid. Vol. 349. Springer International Publishing, pp. 217–232.
Improvement of time series forecasting accuracy is an active research area that has significant importance in many practical domains. Ensemble methods have gained considerable attention from machine learning and soft computing communities in recent years. There are several practical and theoretical reasons, mainly statistical reasons, why an ensemble may be preferred. Ensembles are recognized as one of the most successful approaches to prediction tasks. Previous theoretical studies of ensembles have shown that one of the key reasons for this performance is diversity among ensemble members. Several methods exist to generate diversity. Extensive works in literature suggest that substantial improvements in accuracy can be achieved by combining forecasts from different models. The focus of this chapter will be on ensemble for time series prediction. We describe the use of ensemble methods to compare different models for time series prediction and extensions to the classical ensemble methods for neural networks for classification and regression prediction by using different model architectures. Design, implementation and application will be the main topics of the chapter, and more specifically: conditions under which ensemble based systems may be more beneficial than their single machine; algorithms for generating individual components of ensemble systems; and various procedures through which they can be combined. Various ensemble based algorithms will be analyzed: Bagging, Adaboost and Negative Correlation; as well as combination rules and decision templates. Finally, future directions will be time series forecasting, machine fusion and others areas in which ensemble of machines have shown great promise.

Anghel, D. G. (2021). "Data Snooping Bias in Tests of the Relative Performance of Multiple Forecasting Models." In: *Journal of Banking & Finance* 126, p. 106113.
Tests of the relative performance of multiple forecasting models are sensitive to how the set of alternatives is defined. Evaluating one model against a particular set may show that it has superior predictive ability. However, changing the number or type of alternatives in the set may demonstrate otherwise. This paper focuses on forecasting models based on technical analysis and analyzes how much data snooping bias can occur in tests from restricting the size of forecasting model "universes" or ignoring alternatives used by practitioners and other researchers. A Monte Carlo simulation shows that false discoveries have an average increase of 0.72-2.5 percentage points each time one removes half of the prediction models from the set of relevant alternatives. A complementary empirical investigation suggests that at least 50% of positive findings reported in the literature concerned with trading rule overperformance may be false. Our results motivate several recommendations for applied researchers that would alleviate data snooping bias in some of the more popular statistical tests used in the literature.

Ankile, L. L. and Krange, K. (2022). "The DONUT Approach to Ensemble Combination Forecasting." In: *arXiv e-Print*.
This paper presents an ensemble forecasting method that shows strong results on the M4 Competition dataset by decreasing feature and model selection assumptions, termed DONUT(DO Not UTilize human assumptions). Our assumption reductions, consisting mainly of auto-generated features and a more diverse model pool for the ensemble, significantly outperforms the statistical-feature-based ensemble method FFORMA by Montero-Manso et al. (2020). Furthermore, we investigate feature extraction with a Long short-term memory Network(LSTM) Autoencoder and find that such features contain crucial information not captured by traditional statistical feature approaches. The ensemble weighting model uses both LSTM features and statistical features to combine the models accurately. Analysis of feature importance and interaction show a slight superiority for LSTM features over the statistical ones alone. Clustering analysis shows that different essential LSTM features are different from most statistical features and each other. We also find that increasing the solution space of the weighting model by augmenting the ensemble with new models is something the weighting model learns to use, explaining part

of the accuracy gains. Lastly, we present a formal ex-post-facto analysis of optimal combination and selection for ensembles, quantifying differences through linear optimization on the M4 dataset. We also include a short proof that model combination is superior to model selection, a posteriori.

Ansari, A. F., Benidis, K., Kurle, R., Turkmen, A. C., Soh, H., Smola, A. J., Wang, Y., and Januschowski, T. (2021). "Deep Explicit Duration Switching Models for Time Series." In: *arXiv e-Print*.
Many complex time series can be effectively subdivided into distinct regimes that exhibit persistent dynamics. Discovering the switching behavior and the statistical patterns in these regimes is important for understanding the underlying dynamical system. We propose the Recurrent Explicit Duration Switching Dynamical System (RED-SDS), a flexible model that is capable of identifying both state- and time-dependent switching dynamics. State-dependent switching is enabled by a recurrent state-to-switch connection and an explicit duration count variable is used to improve the time-dependent switching behavior. We demonstrate how to perform efficient inference using a hybrid algorithm that approximates the posterior of the continuous states via an inference network and performs exact inference for the discrete switches and counts. The model is trained by maximizing a Monte Carlo lower bound of the marginal log-likelihood that can be computed efficiently as a byproduct of the inference routine. Empirical results on multiple datasets demonstrate that RED-SDS achieves considerable improvement in time series segmentation and competitive forecasting performance against the state of the art.

Aparicio, D. and Lopez de Prado, M. (2018). "How Hard Is It to Pick the Right Model? MCS and backtest overfitting." In: *Algorithmic Finance* 7, pp. 53–61.
Model selection has become a challenging and pressing need with recent advances in machine learning, artificial intelligence, and the availability of billions of high frequency data signals. However, a majority of model selection methods available in modern finance are subject to backtest overfitting. This is the probability that we select a financial strategy which outperforms during backtest but underperforms in practice. We evaluate the performance of the novel model confidence set (MCS) introduced in Hansen et al. (2011) in a machine learning trading strategy problem. We find that MCS is not robust to multiple testing and that it requires a very high signal-to-noise ratio to be utilizable. More generally, we raise awareness on the use of model selection methods in finance.

Ardia, D. and Dufays, A. (2021). "Measuring uncertainty and uncertainty dispersion from a large set of model predictions." In: *SSRN e-Print*.
We construct measures of uncertainty and its dispersion exploiting the heterogeneity of a large set of model predictions. The approach is forward-looking, can be computed in real-time, and can be applied at any frequency. We illustrate the methodology with expected shortfall predictions of worldwide equity indices generated from 71 risk models. We use the new measures in asset pricing, risk forecasting, and for explaining the aggregate trading volume of S&P 500 firms.

Arias-Calluari, K., Alonso-Marroquin, F., Najafi, M. N., and Harré, M. (2021). "Methods for forecasting the effect of exogenous risks on stock markets." In: *Physica A: Statistical Mechanics and its Applications* 568, p. 125587.
Markets are subjected to both endogenous and exogenous risks that have caused disruptions to financial and economic markets around the globe, leading eventually to fast stock market declines. In the past, markets have recovered after any economic disruption. On this basis, we focus on the outbreak of COVID-19 as a case study of an exogenous risk and analyze its impact on the Standard and Poor's 500 (S&P500) index. We assumed that the S&P500 index reaches a minimum before rising again in the not-too-distant future. Here we present a forecast model of the S&P500 index based on the breaking news and publicly available information. We assumed that the biggest fall of the S&P500 during the COVID-19 outbreak will occur when the largest daily number of deaths was confirmed. We inferred that the peak number of deaths occurs 2-months since the first confirmed case was reported in the USA based on previous COVID-19 situation reports from other countries. We also compare the S&P500 and the DAX market dynamics around the COVID-19 crisis as well as other previous crises, demonstrating that the impact of market news is highly consistent across these multiple market crises. The forecast is a projection of a prediction with stochastic fluctuations described by -gaussian diffusion process with three spatio-temporal regimes. Our forecast was made on the premise that any market response can be decomposed into an overall deterministic trend and a stochastic term. The prediction was based on the deterministic part and for this case study is approximated by the extrapolation of the S&P500 data trend in the initial stages of the outbreak. The stochastic fluctuations have the same structure as the one derived from the past 24 years. A reasonable forecast was achieved with 85% of accuracy.

Arnold, S., Henzi, A., and Ziegel, J. F. (2022). "Sequentially valid tests for forecast calibration." In: *arXiv e-Print*.
Forecasting and forecast evaluation are inherently sequential tasks. Predictions are often issued on a regular basis, such as every hour, day, or month, and their quality is monitored continuously. However, the classical statistical

tools for forecast evaluation are static, in the sense that statistical tests for forecast calibration are only valid if the evaluation period is fixed in advance. Recently, e-values have been introduced as a new, dynamic method for assessing statistical significance. An e-value is a non-negative random variable with expected value at most one under a null hypothesis. Large e-values give evidence against the null hypothesis, and the multiplicative inverse of an e-value is a conservative p-value. E-values are particularly suitable for sequential forecast evaluation, since they naturally lead to statistical tests which are valid under optional stopping. This article proposes e-values for testing probabilistic calibration of forecasts, which is one of the most important notions of calibration. The proposed methods are also more generally applicable for sequential goodness-of-fit testing. We demonstrate that the e-values are competitive in terms of power when compared to extant methods, which do not allow sequential testing. Furthermore, they provide important and useful insights in the evaluation of probabilistic weather forecasts.

Arvanitis, S., Post, T., Potì, V., and Karabati, S. (2021). "Nonparametric tests for Optimal Predictive Ability." In: *International Journal of Forecasting* 37(2), pp. 881–898.

A nonparametric method for comparing multiple forecast models is developed and implemented. The hypothesis of Optimal Predictive Ability generalizes the Superior Predictive Ability hypothesis from a single given loss function to an entire class of loss functions. Distinction is drawn between General Loss functions, Convex Loss functions, and Symmetric Convex Loss functions. The research hypothesis is formulated in terms of moment inequality conditions. The empirical moment conditions are reduced to an exact and finite system of linear inequalities based on piecewise-linear loss functions. The hypothesis can be tested in a statistically consistent way using a blockwise Empirical Likelihood Ratio test statistic. A computationally feasible test procedure computes the test statistic using Convex Optimization methods, and estimates conservative, data-dependent critical values using a majorizing chi-square limit distribution and a moment selection method. An empirical application to inflation forecasting reveals that a very large majority of thousands of forecast models are redundant, leaving predominantly Phillips Curve-type models, when convexity and symmetry are assumed.

Ashouri, M., Hyndman, R. J., and Shmueli, G. (2022). "Fast Forecast Reconciliation Using Linear Models." In: *Journal of Computational and Graphical Statistics*.

Forecasting hierarchical or grouped time series using a reconciliation approach involves two steps: computing base forecasts and reconciling the forecasts. Base forecasts can be computed by popular time series forecasting methods such as exponential smoothing (ETS) and Autoregressive Integrated Moving Average (ARIMA) models. The reconciliation step is a linear process that adjusts the base forecasts to ensure they are coherent. However, using ETS or ARIMA for base forecasts can be computationally challenging when there are a large number of series to forecast, as each model must be numerically optimized for each series. We propose a linear model that avoids this computational problem and handles the forecasting and reconciliation in a single step. The proposed method is very flexible in incorporating external data. We illustrate our approach using a dataset on monthly Australian domestic tourism, as well as a simulated dataset. We compare our approach to reconciliation using ETS and ARIMA, and show that our approach is much faster while providing similar levels of forecast accuracy. Supplementary files for this article are available online. The datasets and R codes used in this article are publicly available at github.com/robjhyndman/linear-hierarchical-forecasting. We have also created a Binder interface for our Australian tourism and Wikipedia examples which allows easily running our code in a browser. The demo folder is reachable from `github.com/mahsaashouri/AUS-Wiki-Binder`.

Athanasopoulos, G., Gamakumara, P., Panagiotelis, A., Hyndman, R. J., and Affan, M. (2019). "Hierarchical Forecasting." In: *Macroeconomic Forecasting in the Era of Big Data*. Springer International Publishing, pp. 689–719.

Accurate forecasts of macroeconomic variables are crucial inputs into the decisions of economic agents and policy makers. Exploiting inherent aggregation structures of such variables, we apply forecast reconciliation methods to generate forecasts that are coherent with the aggregation constraints. We generate both point and probabilistic forecasts for the first time in the macroeconomic setting. Using Australian GDP we show that forecast reconciliation not only returns coherent forecasts but also improves the overall forecast accuracy in both point and probabilistic frameworks.

Athanasopoulos, G. and Kourentzes, N. (2020). *On the evaluation of hierarchical forecasts*. Tech. rep. Monash University.

The aim of this note is to provide a thinking road-map and a practical guide to researchers and practitioners working on hierarchical forecasting problems. Evaluating the performance of hierarchical forecasts comes with new challenges stemming from both the statistical structure of the hierarchy and the application context. We discuss four relevant dimensions for researchers and analysts: the scale and units of time series, the issue of sparsity, the decision context and the importance of multiple evaluation windows. We conclude with a series of practical recommendations.

Atiya, A. F. (2020). "Why does forecast combination work so well?" In: *International Journal of Forecasting* 36(1), pp. 197–200.

Forecast combinations were big winners in the M4 competition. This note reflects on and analyzes the reasons for the success of forecast combination. We illustrate graphically how and in what cases forecast combinations produce good results. We also study the effects of forecast combination on the bias and the variance of the forecast.

Babiak, M. and Barunik, J. (2020). "Deep Learning, Predictability, and Optimal Portfolio Returns." In: *SSRN e-Print*.

We study optimal dynamic portfolio choice of a long-horizon investor who uses deep learning methods to predict equity returns when forming optimal portfolios. The results show statistically and economically significant out-of-sample portfolio benefits of deep learning as measured by high certainty equivalent returns and Sharpe ratios. Return predictability via deep learning generates substantially improved portfolio performance across different subsamples, particularly the recession periods. These gains are robust to including transaction costs, short-selling and borrowing constraints.

Babii, A., Ball, R. T., Ghysels, E., and Striaukas, J. (2021). "Machine Learning Panel Data Regressions with Heavy-tailed Dependent Data: Theory and Application." In: *SSRN e-Print*.

The paper introduces structured machine learning regressions for heavy-tailed dependent panel data potentially sampled at different frequencies. We focus on the sparse-group LASSO regularization. This type of regularization can take advantage of the mixed frequency time series panel data structures and improve the quality of the estimates. We obtain oracle inequalities for the pooled and fixed effects sparse-group LASSO panel data estimators recognizing that financial and economic data can have fat tails. To that end, we leverage on a new Fuk-Nagaev concentration inequality for panel data consisting of heavy-tailed τ-mixing processes.

Bacher, P., Bergsteinsson, H. G., Frolke, L., Sorensen, M. L., Lemos-Vinasco, J., Liisberg, J., Moller, J. K., Nielsen, H. A., and Madsen, H. (2022). "onlineforecast: An R package for adaptive and recursive forecasting." In: *arXiv e-Print*.

Systems that rely on forecasts to make decisions, e.g. control or energy trading systems, require frequent updates of the forecasts. Usually, the forecasts are updated whenever new observations become available, hence in an online setting. We present the R package onlineforecast that provides a generalized setup of data and models for online forecasting. It has functionality for time-adaptive fitting of linear regression-based models. Furthermore, dynamical and non-linear effects can be easily included in the models. The setup is tailored to enable effective use of forecasts as model inputs, e.g. numerical weather forecast. Users can create new models for their particular system applications and run models in an operational online setting. The package also allows users to easily replace parts of the setup, e.g. use kernel or neural network methods for estimation. The package comes with comprehensive vignettes and examples of online forecasting applications in energy systems, but can easily be applied in all fields where online forecasting is used.

Bahrami, A., Shamsuddin, A., and Uylangco, K. (2019). "Are advanced emerging market stock returns predictable? A regime-switching forecast combination approach." In: *Pacific-Basin Finance Journal* 55, pp. 142–160.

Advanced emerging markets (AEMs) transitioning into developed markets experience far-reaching economic and institutional changes. Developing predictive models of stock returns in AEMs involves challenges of parameter instability and model uncertainty. This study uses Markov regime switching (MRS) models to address parameter instability and a combination forecast approach to mitigate model uncertainty. We find that the MRS model better captures the effects of predictor variables on returns compared to models with time-invariant parameters and produces statistically and economically significant return forecasts. Combining return forecasts from different MRS models further improves return predictability in AEMs. Consequently, employing MRS models in conjunction with the combination forecast approach goes a long way to improving forecast accuracy in AEMs.

Bailey, N., Kapetanios, G., and Pesaran, M. H. (2020). "Measurement of Factor Strength: Theory and Practice." In: *SSRN e-Print*.

This paper proposes an estimator of factor strength and establishes its consistency and asymptotic distribution. The proposed estimator is based on the number of statistically significant factor loadings, taking account of the multiple testing problem. We focus on the case where the factors are observed which is of primary interest in many applications in macroeconomics and finance. We also consider using cross section averages as a proxy in the case of unobserved common factors. We face a fundamental factor identification issue when there are more than one unobserved common factors. We investigate the small sample properties of the proposed estimator by means of Monte Carlo experiments under a variety of scenarios. In general, we find that the estimator, and the associated inference, perform well. The test is conservative under the null hypothesis, but, nevertheless, has excellent power properties, especially when the factor strength is sufficiently high. Application of the proposed estimation strategy to factor models of asset returns shows that out of 146 factors recently considered in the finance literature, only the market factor is truly strong, while all other factors are at best semi-strong, with their strength varying considerably over time. Similarly, we only find evidence of semi-strong factors in an updated version of the Stock and Watson (2012) macroeconomic dataset.

Baitinger, E. and Flegel, S. (2021). "New Concepts in Financial Forecasting: Network-Based Information, Topological Data Analysis and their Combination." In: *SSRN e-Print*.

This paper introduces novel financial predictors that are derived from the interaction profile of financial markets. These predictors utilize network-based and topological information. Since these predictors are derived from the inner dynamics (microstructure) of financial markets, they can be best described as microstructural predictors. After equipping the reader with the methodological background of the novel predictors, we perform an extensive in-sample and out-of-sample performance analyses. The in-sample studies demonstrate that microstructural predictors and their combinations are informative with regard to future asset returns. In the out-of-sample studies, we combine microstructual predictors with state of the art machine learning and statistical factor extraction methods. The resulting active forecasting models dominate the benchmark mean model in terms of profitability, but not it terms of statistical precision. Since an investor is usually more concerned with profitability of active investment strategies, the out-of-sample results confirm the value-added of the novel predictors.

Bali, T. G., Goyal, A., Huang, D., Jiang, F., and Wen, Q. (2021). "Different Strokes: Return Predictability Across Stocks and Bonds with Machine Learning and Big Data." In: *SSRN e-Print*.

We investigate the return predictability across stocks and bonds using big data and machine learning. We find that machine learning models substantially improve the out-of-sample performance of stock and bond characteristics in predicting future stock and bond returns. Although both stock and bond characteristics provide strong forecasting power for both stock and bond returns, stock (bond) characteristics do not offer significant incremental predictive power above and beyond bond (stock) characteristics in predicting bond (stock) returns. The results also indicate that stock (bond) characteristics are cash flow (discount rate) predictors and stock (bond) return predictability is driven by mispricing (risk) phenomenon.

Baltas, N. and Karyampas, D. (2020). "Forecasting the Equity Risk Premium: The Importance of Regime-Dependent Evaluation." In: *SSRN e-Print*.

Asset allocation is critically dependent on the ability to forecast the equity risk premium (ERP) out-of-sample. But, is superior econometric predictability across the business cycle synonymous to predictability at all times? We evaluate recently introduced ERP forecasting models, which have been shown to generate econometrically superior ERP forecasts, and find that their forecasting ability is regime-dependent. They give rise to significant relative losses during market downturns, when it matters the most for asset allocators to retain assets and their client base intact. Conversely, any economic benefit occurring during market upswings is diminished for high risk averse and leverage constrained investors.

Baltussen, G., Martens, M., and Penninga, O. (2020). "Predicting Bond Returns: 70 Years of International Evidence." In: *SSRN e-Print*.

We examine the predictability of government bond returns using a deep sample spanning 70 years of international data across the major bond markets. Using an economic, trading-based testing framework we find strong economic and statistical evidence of bond return predictability with a Sharpe ratio of 0.87 since 1950. This finding is robust over markets and time periods, including 30 years of out-of-sample data on international bond markets and a set of nine additional countries. Furthermore, the results are consistent over economic environments, including prolonged periods of rising or falling rates, and is exploitable after transaction costs. The predictability relates to predictability in inflation and economic growth. Overall, government bond premia display predictable dynamics and the timing of international bond market returns offers exploitable opportunities to investors.

Baltussen, G., Martens, M., and Penninga, O. (2021). "Predicting Bond Returns: 70 Years of International Evidence." In: *Financial Analysts Journal* 77(3), pp. 133–155.

We use 70 years of international data from the major bond markets to examine bond return predictability through in-sample and out-of-sample tests. Our results reveal economically strong and statistically significant bond return predictability. This finding is robust over markets and time periods, including 30 years of out-of-sample data, prolonged periods of rising or falling rates, and a dataset of nine additional countries. Furthermore, the results are not explained by market or macroeconomic risks, nor can they be easily attributed to transaction costs or other investment frictions. These results reveal predictable dynamics in government bond returns relevant for academics and practitioners.

Bandara, K., Bergmeir, C., and Smyl, S. (2020). "Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach." In: *Expert Systems with Applications* 140, pp. 112896+.

With the advent of Big Data, nowadays in many applications databases containing large quantities of similar time series are available. Forecasting time series in these domains with traditional univariate forecasting procedures leaves great potentials for producing accurate forecasts untapped. Recurrent neural networks (RNNs), and in particular Long Short Term Memory (LSTM) networks, have proven recently that they are able to outperform state-of-the-art univariate time series forecasting methods in this context, when trained across all available time series. However, if the time series database is heterogeneous, accuracy may degenerate, so that on the way towards fully automatic forecasting methods in this space, a notion of similarity between the time series needs to be built into the methods. To this end, we present a prediction model that can be used with different types of RNN models on subgroups of similar time series, which are identified by time series clustering techniques. We assess our proposed methodology using LSTM networks, a widely popular RNN variant, together with various clustering algorithms, such as kMeans, DBScan, Partition Around Medoids (PAM), and Snob. Our method achieves competitive results on benchmarking datasets under competition evaluation procedures. In particular, in terms of mean sMAPE accuracy it consistently outperforms the baseline LSTM model, and outperforms all other methods on the CIF2016 forecasting competition dataset.

Bandara, K., Hewamalage, H., Liu, Y.-H., Kang, Y., and Bergmeir, C. (2021). "Improving the accuracy of global forecasting models using time series data augmentation." In: *Pattern Recognition* 120, p. 108148.

Forecasting models that are trained across sets of many time series, known as Global Forecasting Models (GFM), have shown recently promising results in forecasting competitions and real-world applications, outperforming many state-of-the-art univariate forecasting techniques. In most cases, GFMs are implemented using deep neural networks, and in particular Recurrent Neural Networks (RNN), which require a sufficient amount of time series to estimate their numerous model parameters. However, many time series databases have only a limited number of time series. In this study, we propose a novel, data augmentation based forecasting framework that is capable of improving the baseline accuracy of the GFM models in less data-abundant settings. We use three time series augmentation techniques: GRATIS, moving block bootstrap (MBB), and dynamic time warping barycentric averaging (DBA) to synthetically generate a collection of time series. The knowledge acquired from these augmented time series is then transferred to the original dataset using two different approaches: the pooled approach and the transfer learning approach. When building GFMs, in the pooled approach, we train a model on the augmented time series alongside the original time series dataset, whereas in the transfer learning approach, we adapt a pre-trained model to the new dataset. In our evaluation on competition and real-world time series datasets, our proposed variants can significantly improve the baseline accuracy of GFM models and outperform state-of-the-art univariate forecasting methods.

Barillas, F. and Shanken, J. (2019). "Real-time Portfolio Choice Implications of Asset Pricing Models." In: *Consortium on Factor Investing Conference*. Vol. 73.

A plethora of asset pricing factors have been proposed in the literature. We study the problem of an investor who is confronted with this "zoo of factors" and wishes to find an optimal portfolio. We propose a Bayesian asset allocation framework that accounts for uncertainty about the correct pricing model. This entails an optimal degree of economic shrinkage that is beneficial for portfolio performance. Under a wide range of beliefs about the extent of mispricing, we find that considering all asset pricing models that can be formed from a given set of factors leads to real-time performance that is superior to that of the sample tangency portfolio. The superiority in out-of-sample performance is even stronger when some of the factors are redundant, as might be the case when a factor has been data mined.

Barras, L. (2019). "A large-scale approach for evaluating asset pricing models." In: *Journal of Financial Economics* 134(3), pp. 549–569.

Recent studies show that the standard test portfolios do not contain sufficient information to discriminate between asset pricing models. To address this issue, we develop a large-scale approach that expands the cross-section to several thousand portfolios. Our novel approach is simple, widely applicable, and allows for formal evaluation/comparison tests. Its benefits are confirmed in empirical tests of CAPM- and characteristic-based

models. While these models are all misspecified, we uncover striking performance differences between them. In particular, the human capital and conditional CAPMs largely outperform the CAPM, which suggests that labor income and time-varying recession risks are primary concerns for investors.

Barrow, D. K. and Crone, S. F. (2016). "A comparison of AdaBoost algorithms for time series forecast combination." In: *International Journal of Forecasting* 32(4), pp. 1103–1119.
Recently, combination algorithms from machine learning classification have been extended to time series regression, most notably seven variants of the popular AdaBoost algorithm. Despite their theoretical promise their empirical accuracy in forecasting has not yet been assessed, either against each other or against any established approaches of forecast combination, model selection, or statistical benchmark algorithms. Also, none of the algorithms have been assessed on a representative set of empirical data, using only few synthetic time series. We remedy this omission by conducting a rigorous empirical evaluation using a representative set of 111 industry time series and a valid and reliable experimental design. We develop a full-factorial design over derived Boosting meta-parameters, creating 42 novel Boosting variants, and create a further 47 novel Boosting variants using research insights from forecast combination. Experiments show that only few Boosting meta-parameters increase accuracy, while meta-parameters derived from forecast combination research outperform others.

Bates, S., Hastie, T., and Tibshirani, R. (2021). "Cross-validation: what does it estimate and how well does it do it?" In: *arXiv e-Print*.
Cross-validation is a widely-used technique to estimate prediction error, but its behavior is complex and not fully understood. Ideally, one would like to think that cross-validation estimates the prediction error for the model at hand, fit to the training data. We prove that this is not the case for the linear model fit by ordinary least squares; rather it estimates the average prediction error of models fit on other unseen training sets drawn from the same population. We further show that this phenomenon occurs for most popular estimates of prediction error, including data splitting, bootstrapping, and Mallow's Cp. Next, the standard confidence intervals for prediction error derived from cross-validation may have coverage far below the desired level. Because each data point is used for both training and testing, there are correlations among the measured accuracies for each fold, and so the usual estimate of variance is too small. We introduce a nested cross-validation scheme to estimate this variance more accurately, and show empirically that this modification leads to intervals with approximately correct coverage in many examples where traditional cross-validation intervals fail. Lastly, our analysis also shows that when producing confidence intervals for prediction accuracy with simple data splitting, one should not re-fit the model on the combined data, since this invalidates the confidence intervals.

Bauer, A., Zufle, M., Herbst, N., Kounev, S., and Curtef, V. (2020). "Telescope: An Automatic Feature Extraction and Transformation Approach for Time Series Forecasting on a Level-Playing Field." In: *Proceedings of the 36th International Conference on Data Engineering (ICDE)*, pp. 1902–1905.
One central problem of machine learning is the inherent limitation to predict only what has been learned - stationarity. Any time series property that eludes stationarity poses a challenge for the proper model building. Furthermore, existing forecasting methods lack reliable forecast accuracy and time-to-result if not applied in their sweet spot. In this paper, we propose a fully automated machine learning-based forecasting approach. Our Telescope approach extracts and transforms features from an input time series and uses them to generate an optimized forecast model. In a broad competition including the latest hybrid forecasters, established statistical, and machine learning-based methods, our Telescope approach shows the best forecast accuracy coupled with a lower and reliable time-to-result.

Bektic, D., Hachenberg, B., and Schiereck, D. (2020). "Factor-based investing in government bond markets: a survey of the current state of research." In: *Journal of Asset Management* 21, pp. 94–105.
Factor investing has become very popular during the last decades, especially with respect to equity markets. After extending Fama-French factors to corporate bond markets, recent research more often concentrates on the government bond space and reveals that there is indeed clear empirical evidence for the existence of significant government bond factors. Voices that state the opposite refer to outdated data samples. By the documentation of rather homogeneous recent empirical evidence, this review underlines the attractiveness of more sophisticated investment approaches, which are well established in equity and even in corporate bond markets, to the segment of government bonds.

Ben Baccar, Y. (2019). "Comparative Study on Time Series Forecasting Models." en. MA thesis. ParisTech.
Today, there are plenty of various forecasting models for Time Series with each one requiring proper data preprocessing and analysis to provide a usable prediction. The aim of this report is to conduct a comparative study on the most commonly used Time Series estimators in order to benchmark their performance on a wide

variety of series from different fields (economics, finance, meteorology, etc...) and compare them to the in-house estimator developed by SAP called SAPTF. All the implemented models are automated, making hyper-parameter search a part of the model, this is done so that it can be used without any prior knowledge of the models, nor the datasets on which they will be applied on.

Benidis, K., Rangapuram, S. S., Flunkert, V., Wang, B., Maddix, D., Turkmen, C., Gasthaus, J., Bohlke-Schneider, M., Salinas, D., Stella, L., Callot, L., and Januschowski, T. (2020). "Neural forecasting: Introduction and literature overview." In: *arXiv e-Print*.

Neural network based forecasting methods have become ubiquitous in large-scale industrial forecasting applications over the last years. As the prevalence of neural network based solutions among the best entries in the recent M4 competition shows, the recent popularity of neural forecasting methods is not limited to industry and has also reached academia. This article aims at providing an introduction and an overview of some of the advances that have permitted the resurgence of neural networks in machine learning. Building on these foundations, the article then gives an overview of the recent literature on neural networks for forecasting and applications.

Berardi, A., Markovich, M., Plazzi, A., and Tamoni, A. (2021). "Mind the (Convergence) Gap: Bond Predictability Strikes Back!" In: *Management Science*.

We show that the difference between the natural rate of interest and the current level of monetary policy stance, which we label Convergence Gap (CG), contains information that is valuable for bond predictability. Adding CG in forecasting regressions of bond excess returns significantly raises the R2, and restores countercyclical variation in bond risk premia that is otherwise missed by forward rates. Consistent with the argument that CG captures the effect of real imbalances on the path of rates, our factor has predictive ability for real bond excess returns. The importance of the gap remains robust out-of-sample and in countries other than the United States. Furthermore, its inclusion brings significant economic gains in the context of dynamic conditional asset allocation.

Bergmeir, C., Triguero, I., Molina, D., Aznarte, J. L., and Benitez, J. M. (2012). "Time Series Modeling and Forecasting Using Memetic Algorithms for Regime-Switching Models." In: *IEEE Transactions on Neural Networks and Learning Systems* 23(11), pp. 1841–1847.

In this brief, we present a novel model fitting procedure for the neuro-coefficient smooth transition autoregressive model (NCSTAR), as presented by Medeiros and Veiga. The model is endowed with a statistically founded iterative building procedure and can be interpreted in terms of fuzzy rule-based systems. The interpretability of the generated models and a mathematically sound building procedure are two very important properties of forecasting models. The model fitting procedure employed by the original NCSTAR is a combination of initial parameter estimation by a grid search procedure with a traditional local search algorithm. We propose a different fitting procedure, using a memetic algorithm, in order to obtain more accurate models. An empirical evaluation of the method is performed, applying it to various real-world time series originating from three forecasting competitions. The results indicate that we can significantly enhance the accuracy of the models, making them competitive to models commonly used in the field.

Bergmeir, C., Hyndman, R. J., and Koo, B. (2018). "A note on the validity of cross-validation for evaluating autoregressive time series prediction." In: *Computational Statistics & Data Analysis* 120, pp. 70–83.

One of the most widely used standard procedures for model evaluation in classification and regression is K fold cross-validation (CV). However, when it comes to time series forecasting, because of the inherent serial correlation and potential non-stationarity of the data, its application is not straightforward and often replaced by practitioners in favour of an out-of-sample (OOS) evaluation. It is shown that for purely autoregressive models, the use of standard K fold CV is possible provided the models considered have uncorrelated errors. Such a setup occurs, for example, when the models nest a more appropriate model. This is very common when Machine Learning methods are used for prediction, and where CV can control for overfitting the data. Theoretical insights supporting these arguments are presented, along with a simulation study and a real-world example. It is shown empirically that K fold CV performs favourably compared to both OOS evaluation and other time-series-specific techniques such as non-dependent cross-validation.

Bernardi, M. and Catania, L. (2014). "The Model Confidence Set package for R." In: *arXiv e-Print*.

This paper presents the R package MCS which implements the Model Confidence Set (MCS) procedure recently developed by Hansen et al. (2011). The Hansen's procedure consists on a sequence of tests which permits to construct a set of 'superior' models, where the null hypothesis of Equal Predictive Ability (EPA) is not rejected at a certain confidence level. The EPA statistic tests is calculated for an arbitrary loss function, meaning that we could test models on various aspects, for example punctual forecasts. The relevance of the package is shown using an example which aims at illustrating in details the use of the functions provided by the package. The example compares the ability of different models belonging to the ARCH family to predict large financial losses. We also discuss the implementation of the ARCH–type models and their maximum likelihood estimation using the popular R package rugarch developed by Ghalanos (2014).

Berrisch, J. and Ziel, F. (2021). "CRPS Learning." In: *arXiv e-Print*.

Combination and aggregation techniques can significantly improve forecast accuracy. This also holds for probabilistic forecasting methods where predictive distributions are combined. There are several time-varying and adaptive weighting schemes such as Bayesian model averaging (BMA). However, the quality of different forecasts may vary not only over time but also within the distribution. For example, some distribution forecasts may be more accurate in the center of the distributions, while others are better at predicting the tails. Therefore, we introduce a new weighting method that considers the differences in performance over time and within the distribution. We discuss pointwise combination algorithms based on aggregation across quantiles that optimize with respect to the continuous ranked probability score (CRPS). After analyzing the theoretical properties of pointwise CRPS learning, we discuss B- and P-Spline-based estimation techniques for batch and online learning, based on quantile regression and prediction with expert advice. We prove that the proposed fully adaptive Bernstein online aggregation (BOA) method for pointwise CRPS online learning has optimal convergence properties. They are confirmed in simulations and a probabilistic forecasting study for European emission allowance (EUA) prices.

Bessembinder, H. (, Burt, A. P., and Hrdlicka, C. M. (2022). "Time Series Variation in the Factor Zoo." In: *SSRN e-Print*.

Should we be surprised at the number of "animals" in the "Factor Zoo"? The ability of the CAPM as well as workhorse three- to six-factor models to explain the cross-section of returns varies substantially over time, providing scope for a broad set of factors. We study 205 previously-identified factors, documenting time variation in their significance, prior to, during, and after the periods studied by the original authors. The number of statistically significant factors, as well as the number of principal components obtained from them, varies with the cross-sectional dispersion in individual stock CAPM alphas. The number of significant factors is strongly related to the number of publicly-listed firms, and is also related to institutional ownership, recession indicators, interest rates, and measures of diversity in firm characteristics. On balance, our results suggest that the large number of factors with significant explanatory power reflects the complexity of the economic environment, including changes in investor composition, the types of firms listed, and competitive conditions.

Bhatnagar, A., Kassianik, P., Liu, C., Lan, T., Yang, W., Cassius, R., Sahoo, D., Arpit, D., Subramanian, S., Woo, G., Saha, A., Jagota, A. K., Gopalakrishnan, G., Singh, M., Krithika, K. C., Maddineni, S., Cho, D., Zong, B., Zhou, Y., Xiong, C., Savarese, S., Hoi, S., and Wang, H. (2021). "Merlion: A Machine Learning Library for Time Series." In: *arXiv e-Print*.

We introduce Merlion, an open-source machine learning library for time series. It features a unified interface for many commonly used models and datasets for anomaly detection and forecasting on both univariate and multivariate time series, along with standard pre/post-processing layers. It has several modules to improve ease-of-use, including visualization, anomaly score calibration to improve interpetability, AutoML for hyperparameter tuning and model selection, and model ensembling. Merlion also provides a unique evaluation framework that simulates the live deployment and re-training of a model in production. This library aims to provide engineers and researchers a one-stop solution to rapidly develop models for their specific time series needs and benchmark them across multiple time series datasets. In this technical report, we highlight Merlion's architecture and major functionalities, and we report benchmark numbers across different baseline models and ensembles.

Bianchi, D., Buchner, M., and Tamoni, A. (2021). "Bond Risk Premiums with Machine Learning." In: *The Review of Financial Studies* 34(2), pp. 1046–1089.

We show that machine learning methods, in particular, extreme trees and neural networks (NNs), provide strong statistical evidence in favor of bond return predictability. NN forecasts based on macroeconomic and yield information translate into economic gains that are larger than those obtained using yields alone. Interestingly, the nature of unspanned factors changes along the yield curve: stock- and labor-market-related variables are more relevant for short-term maturities, whereas output and income variables matter more for longer maturities. Finally, NN forecasts correlate with proxies for time-varying risk aversion and uncertainty, lending support to models featuring both channels.

Bianchi, D. and McAlinn, K. (2021). "Divide and Conquer: Financial Ratios and Industry Returns Predictability." In: *SSRN e-Print*.

We propose a novel approach for forecasting the equity premium within a data-rich environment based on ensembling small-scale linear models. The economic nature of the predictors is exploited to efficiently retain all of the information available without assuming a priori that some predictor might be irrelevant or easily reducible to a latent factor. Empirically, our results lend strong support for transparent linear predictive models and the use of accounting-based information when forecasting both industry and aggregate stock market excess returns: positive statistical and economic out-of-sample performance compared to sparse predictive regressions, forecast combination strategies and complex non-linear machine learning algorithms.

Bianchi, D. and Tamoni, A. (2020). "Sparse Predictive Regressions: Statistical Performance and Economic Significance." In: *Machine Learning for Asset Management: New Developments and Financial Applications*. Ed. by E. Jurczenko. Wiley, pp. 75–113.

This chapter proposes and evaluates a variety of penalized regression methods for forecasting and economic decision-making in a data-rich environment under parameter uncertainty. Empirically, it explores the statistical and economic performance across different asset classes such as stocks, bonds and currencies, and alternative strategies within an asset class. The main results show that penalty terms that shrink the model space and regularize the remaining regression coefficients tend to outperform both dense (e.g. ridge regression) and sparse (e.g. lasso regression) methodologies, as well as that the amount of shrinkage tends to evolve over time and differs across asset classes. The chapter discusses the use of machine learning in financial economics and describes the data with particular emphasis on the various asset classes. It also provides an overview of the shrinkage priors adopted in the analysis of statistical and economic performance of asset classes.

Bielinski, A. and Broby, D. (2021). "Machine Learning Methods in Asset Pricing." In: *SSRN e-Print*.

This paper evaluates the traditional asset pricing models and examines the literature on the most promising machine learning techniques that can be used to price securities. Asset price forecasting is essential to efficient markets. Capital Asset Pricing Models (CAPM), Arbitrage Pricing Theory (APT) and a multitude of Factor Models are used to price securities and to establish mean variance optimal portfolios. An increasing number of scholars and financial practitioners have begun to explore the role of machine learning in asset pricing. We show how these methods have been applied in academia and discuss their results in maximizing the Sharpe Ratio. We also explore the potential use of neural networks in asset pricing. We believe that their capacity to process large amounts of data and their ability to accurately capture non-linear relationships makes them a useful estimation tool.

Billio, M., Casarin, R., Ravazzolo, F., and Dijk, H. K. van (2013). "Time-varying combinations of predictive densities using nonlinear filtering." In: *Journal of Econometrics* 177(2), pp. 213–232.

We propose a Bayesian combination approach for multivariate predictive densities which relies upon a distributional state space representation of the combination weights. Several specifications of multivariate time-varying weights are introduced with a particular focus on weight dynamics driven by the past performance of the predictive densities and the use of learning mechanisms. In the proposed approach the model set can be incomplete, meaning that all models can be individually misspecified. A Sequential Monte Carlo method is proposed to approximate the filtering and predictive densities. The combination approach is assessed using statistical and utility-based performance measures for evaluating density forecasts of simulated data, US macroeconomic time series and surveys of stock market prices. Simulation results indicate that, for a set of linear autoregressive models, the combination strategy is successful in selecting, with probability close to one, the true model when the model set is complete and it is able to detect parameter instability when the model set includes the true model that has generated subsamples of data. Also, substantial uncertainty appears in the weights when predictors are similar; residual uncertainty reduces when the model set is complete; and learning reduces this uncertainty. For the macro series we find that incompleteness of the models is relatively large in the 1970's, the beginning of the 1980's and during the recent financial crisis, and lower during the Great Moderation; the predicted probabilities of recession accurately compare with the NBER business cycle dating; model weights have substantial uncertainty attached. With respect to returns of the SandP 500 series, we find that an investment strategy using a combination of predictions from professional forecasters and from a white noise model puts more weight on the white noise model in the beginning of the 1990's and switches to giving more weight to the professional forecasts over time. Information on the complete predictive distribution and not just on some moments turns out to be very important, above all during turbulent times such as the recent financial crisis. More generally, the proposed distributional state space representation offers great flexibility in combining densities.

Bisaglia, L. and Grigoletto, M. (2021). "A new time-varying model for forecasting long-memory series." In: *Statistical Methods & Applications* 30, pp. 139–155.

In this work we propose a new class of long-memory models with time-varying fractional parameter. In particular, the dynamics of the long-memory coefficient, d, is specified through a stochastic recurrence equation driven by the score of the predictive likelihood, as suggested by Creal et al. (J Appl Econom 28:777-795, 2013) and Harvey (Dynamic models for volatility and heavy tails: with applications to financial and economic time series, Cambridge University Press, Cambridge, 2013). We demonstrate the validity of the proposed model by a Monte Carlo experiment and an application to two real time series.

Bjerregård, M. B., Møller, J. K., and Madsen, H. (2021). "An introduction to multivariate probabilistic forecast evaluation." In: *Energy and AI* 4, p. 100058.

Probabilistic forecasting is becoming increasingly important for a wide range of applications, especially for energy systems such as forecasting wind power production. A need for proper evaluation of probabilistic forecasts follows naturally with this, because evaluation is the key to improving the forecasts. Although plenty of excellent reviews and research papers on probabilistic forecast evaluation already exist, we find that there is a need for an introduction with some practical application. In particular, many forecast scenarios in energy systems are inherently multivariate, and while univariate evaluation methods are well understood and documented, only limited and scattered work has been done on their multivariate counterparts. This paper therefore contains a review of a selected set of probabilistic forecast evaluation methods, primarily scoring rules, as well as practical sections that explain how these methods can be calculated and estimated. In three case studies featuring simple autoregressive models, stochastic differential equations and real wind power data, we implement, apply and discuss the logarithmic score, the continuous ranked probability score and the variogram score for forecasting problems of varying dimension. Finally, the advantages and disadvantages of the three scoring rules are highlighted, and this provides a significant step towards deciding on an evaluation method for a given multivariate forecast scenario including forecast scenarios relevant for energy systems.

Blitz, D., Hanauer, M. X., Vidojevic, M., and Vliet, P. v. (2018). "Five Concerns with the Five-Factor Model." In: *The Journal of Portfolio Management* 44(4), pp. 71–78.

The new Fama-French five-factor model is likely to become the new benchmark for asset pricing studies. Although the five-factor model exhibits significantly improved explanatory power compared to its predecessor, the classic three-factor model, the authors identify five concerns with regard to the new model. First, it maintains the capital asset pricing model relation between market beta and return, despite mounting evidence that the empirical relation is flat, or even negative. Second, it continues to ignore the, by now widely accepted, momentum effect. Third, there are a number of robustness concerns with regard to the two new factors, profitability and investment. Fourth, whereas risk-based explanations were key for justifying the factors in the three-factor model, the economic rationale for the two new factors is much less clear. Fifth and finally, it does not seem likely that the five-factor model is going to settle the main asset pricing debates or lead to consensus.

Bloemheuvel, S., van den Hoogen, J., Jozinovic, D., Michelini, A., and Atzmueller, M. (2022). "Multivariate Time Series Regression with Graph Neural Networks." In: *arXiv e-Print*.

Machine learning, with its advances in Deep Learning has shown great potential in analysing time series in the past. However, in many scenarios, additional information is available that can potentially improve predictions, by incorporating it into the learning methods. This is crucial for data that arises from e.g., sensor networks that contain information about sensor locations. Then, such spatial information can be exploited by modeling it via graph structures, along with the sequential (time) information. Recent advances in adapting Deep Learning to graphs have shown promising potential in various graph-related tasks. However, these methods have not been adapted for time series related tasks to a great extent. Specifically, most attempts have essentially consolidated around Spatial-Temporal Graph Neural Networks for time series forecasting with small sequence lengths. Generally, these architectures are not suited for regression or classification tasks that contain large sequences of data. Therefore, in this work, we propose an architecture capable of processing these long sequences in a multivariate time series regression task, using the benefits of Graph Neural Networks to improve predictions. Our model is tested on two seismic datasets that contain earthquake waveforms, where the goal is to predict intensity measurements of ground shaking at a set of stations. Our findings demonstrate promising results of our approach, which are discussed in depth with an additional ablation study.

Bokde, N. D., Yaseen, Z. M., and Andersen, G. B. (2020). "ForecastTB - An R Package as a Test-Bench for Time Series Forecasting, with Application of Wind Speed and Solar Radiation Modeling." In: *Energies* 13(10), p. 2578.

This paper introduces an R package ForecastTB that can be used to compare the accuracy of different forecasting methods as related to the characteristics of a time series dataset. The ForecastTB is a plug-and-play structured module, and several forecasting methods can be included with simple instructions. The proposed test-bench is not limited to the default forecasting and error metric functions, and users are able to append, remove, or choose the desired methods as per requirements. Besides, several plotting functions and statistical performance metrics are provided to visualize the comparative performance and accuracy of different forecasting methods. Furthermore, this paper presents real application examples with natural time series datasets (i.e., wind speed and solar radiation) to exhibit the features of the ForecastTB package to evaluate forecasting comparison analysis as affected by the characteristics of a dataset. Modeling results indicated the applicability and robustness of the proposed R package ForecastTB for time series forecasting.

Botchkarev, A. (2019). "A new typology design of performance metrics to measure errors in machine learning regression algorithms." In: *Interdisciplinary Journal of Information, Knowledge, and Management* 14, pp. 045–076.

Aim/Purpose: The aim of this study was to analyze various performance metrics and approaches to their classification. The main goal of the study was to develop a new typology that will help to advance knowledge of metrics and facilitate their use in machine learning regression algorithms Background: Performance metrics (error measures) are vital components of the evaluation frameworks in various fields. A performance metric can be defined as a logical and mathematical construct designed to measure how close are the actual results from what has been expected or predicted. A vast variety of performance metrics have been described in academic literature. The most commonly mentioned metrics in research studies are Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), etc. Knowledge about metrics properties needs to be systematized to simplify the design and use of the metrics. Methodology: A qualitative study was conducted to achieve the objectives of identifying related peer-reviewed research studies, literature reviews, critical thinking and inductive reasoning. Contribution: The main contribution of this paper is in ordering knowledge of performance metrics and enhancing understanding of their structure and properties by proposing a new typology, generic primary metrics mathematical formula and a visualization chart Findings: Based on the analysis of the structure of numerous performance metrics, we proposed a framework of metrics which includes four (4) categories: primary metrics, extended metrics, composite metrics, and hybrid sets of metrics. The paper identified three (3) key components (dimensions) that determine the structure and properties of primary metrics: method of determining point distance, method of normalization, method of aggregation of point distances over a data set. For each component, implementation options have been identified. The suggested new typology has been shown to cover a total of over 40 commonly used primary metrics Recommendations for Practitioners: Presented findings can be used to facilitate teaching performance metrics to university students and expedite metrics selection and implementation processes for practitioners Recommendation for Researchers: By using the proposed typology, researchers can streamline development of new metrics with predetermined properties Impact on Society: The outcomes of this study could be used for improving evaluation results in machine learning regression, forecasting and prognostics with direct or indirect positive impacts on innovation and productivity in a societal sense Future Research: Future research is needed to examine the properties of the extended metrics, composite metrics, and hybrid sets of metrics. Empirical study of the metrics is needed using R Studio or Azure Machine Learning Studio, to find associations between the properties of primary metrics and their behavior in a wide spectrum of data characteristics and business or research requirements.

Bouallegue, Z. B., Haiden, T., and Richardson, D. S. (2018). "The diagonal score: Definition, properties, and interpretations." In: *Quarterly Journal of the Royal Meteorological Society* 144(714), pp. 1463–1473.

The diagonal score is proposed as a new scoring rule for the assessment of univariate probabilistic forecasts. First, a general framework for the definition of proper user-oriented scores is set up using the concept of elementary scoring rules. Building on this fundamental notion, a complete overview of forecast skill is provided by a new verification tool. The forecast skill card introduced displays the forecast value as a function of two decision parameters, a probability level and an event threshold, simultaneously. The diagonal score emerges as a summary score focusing on the diagonal of the skill card, that is, by fixing the relationship between the event base rate and forecast probability level. The properties of the new score, as well as interpretations from the perspectives of users and developers, are discussed based on synthetic datasets and 2 m temperature forecasts of the operational ECMWF ensemble prediction system.

Brehmer, J., Gneiting, T., Schlather, M., and Strokorb, K. (2021). "Using scoring functions to evaluate point process forecasts." In: *arXiv e-Print*.

Point process models are widely used tools to issue forecasts or assess risks. In order to check which models are useful in practice, they are examined by a variety of statistical methods. We transfer the concept of consistent scoring functions, which are principled statistical tools to compare forecasts, to the point process setting. The results provide a novel approach for the comparative assessment of forecasts and models and encompass some existing testing procedures.

Brehmer, J. R. and Gneiting, T. (2020). "Properization: constructing proper scoring rules via Bayes acts." In: *Annals of the Institute of Statistical Mathematics* 72(3), pp. 659–673.
Scoring rules serve to quantify predictive performance. A scoring rule is proper if truth telling is an optimal strategy in expectation. Subject to customary regularity conditions, every scoring rule can be made proper, by applying a special case of the Bayes act construction studied by Grunwald and Dawid (Ann Stat 32:1367-1433, 2004) and Dawid (Ann Inst Stat Math 59:77-93, 2007), to which we refer as properization. We discuss examples from the recent literature and apply the construction to create new types, and reinterpret existing forms, of proper scoring rules and consistent scoring functions. In an abstract setting, we formulate sufficient conditions under which Bayes acts exist and scoring rules can be made proper.

Breitung, J. and Knuppel, M. (2021). "How far can we forecast? Statistical tests of the predictive content." In: *Journal of Applied Econometrics* 36(4), pp. 369–392.
We develop tests for the null hypothesis that forecasts become uninformative beyond some maximum forecast horizon. The forecast may result from a survey of forecasters or from an estimated parametric model. The first class of tests compares the mean-squared prediction error of the forecast to the variance of the evaluation sample, whereas the second class of tests compares it with the mean-squared prediction error of the recursive mean. We show that the forecast comparison may easily be performed by adopting the encompassing principle, which results in simple regression tests with standard asymptotic inference. Our tests are applied to forecasts of macroeconomic key variables from the survey of Consensus Economics. The results suggest that these forecasts are barely informative beyond two to four quarters ahead.

Bryzgalova, S., Huang, J., and Julliard, C. (2021). "Bayesian solutions for the factor zoo: we just ran two quadrillion models." In: *SSRN e-Print*.
We propose a novel, and simple, Bayesian estimation and model selection procedure for cross-sectional asset pricing. Our approach, that allows for both tradable and non-tradable factors, and is applicable to high dimensional cases, has several desirable properties. First, weak and spurious factors lead to diffuse, and centered at zero, posteriors for their market price of risk, making such factors easily detectable. Second, posterior inference is robust to the presence of such factors. Third, we show that flat priors for risk premia lead to improper marginal likelihoods, rendering model selection invalid. Therefore, we provide a novel prior, that is diffuse for strong factors but shrinks away useless ones, under which posterior probabilities are well behaved, and can be used for factor and (non necessarily nested) model selection, as well as model averaging, in large scale problems. We apply our method to a very large set of factors proposed in the literature, and analyse 2.25 quadrillion possible models, gaining novel insights on the empirical drivers of asset returns.

Bulut, L. (2019). "Does Statistical Significance Help to Evaluate Predictive Performance of Competing Models?" In: *SSRN e-Print*.
In Monte Carlo experiment with simulated data, we show that as a point forecast criterion, the Clark and West's (2006) unconditional test of mean squared prediction errors does not reflect the relative performance of a superior model over a relatively weaker one. The simulation results show that even though the mean squared prediction errors of a constructed superior model is far below a weaker alternative, the Clark- West test does not reflect this in their test statistics. Therefore, studies that use this statistic in testing the predictive accuracy of alternative exchange rate models, stock return predictability, inflation forecasting, and unemployment forecasting should not weight too much on the magnitude of the statistically significant Clark-West tests statistics.

Burns, D. M. and Whyne, C. M. (2018). "Seglearn: A Python Package for Learning Sequences and Time Series." In: *arXiv e-Print*.
Seglearn is an open-source python package for machine learning time series or sequences using a sliding window segmentation approach. The implementation provides a flexible pipeline for tackling classification, regression, and forecasting problems with multivariate sequence and contextual data. This package is compatible with scikit-learn and is listed under scikit-learn Related Projects. The package depends on numpy, scipy, and scikit-learn. Seglearn is distributed under the BSD 3-Clause License. Documentation includes a detailed API description, user guide, and examples. Unit tests provide a high degree of code coverage.

Cai, Z., Fang, Y., and Xu, Q. (2022). "Testing capital asset pricing models using functional-coefficient panel data models with cross-sectional dependence." In: *Journal of Econometrics* 227(1), pp. 114–133.
This paper proposes a functional-coefficient panel data model with cross-sectional dependence motivated by re-examining the empirical performance of conditional capital asset pricing model. In order to characterize

the time-varying property of assets' betas and alpha, our proposed model allows the betas to be unknown functions of some macroeconomic and financial instruments. Moreover, a common factor structure is introduced to characterize cross-sectional dependence which is an attractive feature under a panel data regression setting as different assets or portfolios may be affected by same unobserved shocks. Compared to the existing studies, such as the classic Fama-MacBeth two-step procedure, our model can achieve substantial efficiency gains for inference by adopting a one-step procedure using the entire sample rather than a single cross-sectional regression at each time point. We propose a local linear common correlated effects estimator for estimating time-varying betas by pooling the data. The consistency and asymptotic normality of the proposed estimators are established. Another methodological and empirical challenge in asset pricing is how to test the constancy of conditional betas and the significance of pricing errors, we echo this challenge by constructing an -norm statistic for functional-coefficient panel data models allowing for cross-sectional dependence. We show that the new test statistic has a limiting standard normal distribution under the null hypothesis. Finally, the method is applied to test the model in Fama and French (1993) using Fama-French 25 and 100 portfolios, sorted by size and book-to-market ratio, respectively, dated from July 1963 to July 2018.

Caldeira, J. F., Moura, G. V., and Santos, A. A. P. (2018). "Yield curve forecast combinations based on bond portfolio performance." In: *Journal of Forecasting* 37(1), pp. 64–82.

We propose an economically motivated forecast combination strategy in which model weights are related to portfolio returns obtained by a given forecast model. An empirical application based on an optimal mean-variance bond portfolio problem is used to highlight the advantages of the proposed approach with respect to combination methods based on statistical measures of forecast accuracy. We compute average net excess returns, standard deviation, and the Sharpe ratio of bond portfolios obtained with nine alternative yield curve specifications, as well as with 12 different forecast combination strategies. Return-based forecast combination schemes clearly outperformed approaches based on statistical measures of forecast accuracy in terms of economic criteria. Moreover, return-based approaches that dynamically select only the model with highest weight each period and discard all other models delivered even better results, evidencing not only the advantages of trimming forecast combinations but also the ability of the proposed approach to detect best-performing models. To analyze the robustness of our results, different levels of risk aversion and a different dataset are considered.

Capolongo, A. and Pacella, C. (2021). "Forecasting inflation in the euro area: countries matter!" In: *Empirical Economics* 61, pp. 2477–2499.

We construct a Bayesian vector autoregressive model with three layers of information: the key drivers of inflation, cross-country dynamic interactions, and country-specific variables. The model provides good forecasting accuracy with respect to the popular benchmarks used in the literature. We perform a step-by-step analysis to shed light on which layer of information is more crucial for accurately forecasting medium-run euro area inflation. Our empirical analysis reveals the importance of including the key drivers of inflation and taking into account the multi-country dimension of the euro area. The results show that the complete model performs better overall in forecasting inflation excluding energy and unprocessed food over the medium term. We use the model to establish stylized facts on the euro area and cross-country heterogeneity over the business cycle.

Carr, P. P. and Wu, L. (2021). *Decomposing Long Bond Returns: A Decentralized Theory*. Tech. rep. NYU.

Classic bond pricing links, i.e., centralizes, bond valuation across all maturities by specifying the dynamics of the short-term interest rate. This paper develops a decentralized theory that prices each bond based purely on the near-term behavior of the bond's own yield. The theory levers the domain expertise of an investor on a particular bond and allows the investor to make pricing and investment analysis on the bond without the shackles of an ambitious centralizing mandate. The theory decomposes the short-term return on a bond with respect to the variation of its own yield. Imposing no dynamic arbitrage on the return decomposition leads to a simple pricing equation relating the bond yield to the market pricing and conditional mean and variance forecasts of the yield's near-term change. The paper illustrates the theory's applications in decentralized investment of a single bond and in the construction and investment of decentralized butterfly bond portfolios.

Castilho, D., Souza, T. T. P., Kang, S. M., Gama, J., and Carvalho, A. C. P. L. F. de (2021). "Forecasting Financial Market Structure from Network Features using Machine Learning." In: *arXiv e-Print*.

We propose a model that forecasts market correlation structure from link- and node-based financial network features using machine learning. For such, market structure is modeled as a dynamic asset network by quantifying time-dependent co-movement of asset price returns across company constituents of major global market indices. We provide empirical evidence using three different network filtering methods to estimate market structure, namely Dynamic Asset Graph (DAG), Dynamic Minimal Spanning Tree (DMST) and Dynamic Thresh-

old Networks (DTN). Experimental results show that the proposed model can forecast market structure with high predictive performance with up to 40% improvement over a time-invariant correlation-based benchmark. Non-pair-wise correlation features showed to be important compared to traditionally used pair-wise correlation measures for all markets studied, particularly in the long-term forecasting of stock market structure. Evidence is provided for stock constituents of the DAX30, EUROSTOXX50, FTSE100, HANGSENG50, NASDAQ100 and NIFTY50 market indices. Findings can be useful to improve portfolio selection and risk management methods, which commonly rely on a backward-looking covariance matrix to estimate portfolio risk.

Castle, J. L., Doornik, J. A., and Hendry, D. F. (2021a). "Forecasting Principles from Experience with Forecasting Competitions." In: *Forecasting* 3(1), pp. 138–165.
Economic forecasting is difficult, largely because of the many sources of nonstationarity influencing observational time series. Forecasting competitions aim to improve the practice of economic forecasting by providing very large data sets on which the efficacy of forecasting methods can be evaluated. We consider the general principles that seem to be the foundation for successful forecasting, and show how these are relevant for methods that did well in the M4 competition. We establish some general properties of the M4 data set, which we use to improve the basic benchmark methods, as well as the Card method that we created for our submission to that competition. A data generation process is proposed that captures the salient features of the annual data in M4.

Castle, J. L., Doornik, J. A., and Hendry, D. F. (2021b). "Selecting a Model for Forecasting." In: *Econometrics* 9(3), p. 26.
We investigate forecasting in models that condition on variables for which future values are unknown. We consider the role of the significance level because it guides the binary decisions whether to include or exclude variables. The analysis is extended by allowing for a structural break, either in the first forecast period or just before. Theoretical results are derived for a three-variable static model, but generalized to include dynamics and many more variables in the simulation experiment. The results show that the trade-off for selecting variables in forecasting models in a stationary world, namely that variables should be retained if their noncentralities exceed unity, still applies in settings with structural breaks. This provides support for model selection at looser than conventional settings, albeit with many additional features explaining the forecast performance, and with the caveat that retaining irrelevant variables that are subject to location shifts can worsen forecast performance.

Cerqueira, V., Torgo, L., and Mozetič, I. (2020). "Evaluating time series forecasting models: an empirical study on performance estimation methods." In: *Machine Learning* 109, pp. 1997–2028.
Performance estimation aims at estimating the loss that a predictive model will incur on unseen data. This process is a fundamental stage in any machine learning project. In this paper we study the application of these methods to time series forecasting tasks. For independent and identically distributed data the most common approach is cross-validation. However, the dependency among observations in time series raises some caveats about the most appropriate way to estimate performance in this type of data. Currently, there is no consensual approach. We contribute to the literature by presenting an extensive empirical study which compares different performance estimation methods for time series forecasting tasks. These methods include variants of cross-validation, out-of-sample (holdout), and prequential approaches. Two case studies are analysed: One with 174 real-world time series and another with three synthetic time series. Results show noticeable differences in the performance estimation methods in the two scenarios. In particular, empirical experiments suggest that blocked cross-validation can be applied to stationary time series. However, when the time series are non-stationary, the most accurate estimates are produced by out-of-sample methods, particularly the holdout approach repeated in multiple testing periods.

Cerqueira, V., Torgo, L., and Soares, C. (2021a). "Model Selection for Time Series Forecasting: Empirical Analysis of Different Estimators." In: *arXiv e-Print*.
Evaluating predictive models is a crucial task in predictive analytics. This process is especially challenging with time series data where the observations show temporal dependencies. Several studies have analysed how different performance estimation methods compare with each other for approximating the true loss incurred by a given forecasting model. However, these studies do not address how the estimators behave for model selection: the ability to select the best solution among a set of alternatives. We address this issue and compare a set of estimation methods for model selection in time series forecasting tasks. We attempt to answer two main questions: (i) how often is the best possible model selected by the estimators; and (ii) what is the performance loss when it does not. We empirically found that the accuracy of the estimators for selecting the best solution is low, and the overall forecasting performance loss associated with the model selection process ranges from 1.2%

to 2.3%. We also discovered that some factors, such as the sample size, are important in the relative performance of the estimators.

Cerqueira, V., Torgo, L., Soares, C., and Bifet, A. (2021b). "Model Compression for Dynamic Forecast Combination." In: *arXiv e-Print*.
The predictive advantage of combining several different predictive models is widely accepted. Particularly in time series forecasting problems, this combination is often dynamic to cope with potential non-stationary sources of variation present in the data. Despite their superior predictive performance, ensemble methods entail two main limitations: high computational costs and lack of transparency. These issues often preclude the deployment of such approaches, in favour of simpler yet more efficient and reliable ones. In this paper, we leverage the idea of model compression to address this problem in time series forecasting tasks. Model compression approaches have been mostly unexplored for forecasting. Their application in time series is challenging due to the evolving nature of the data. Further, while the literature focuses on neural networks, we apply model compression to distinct types of methods. In an extensive set of experiments, we show that compressing dynamic forecasting ensembles into an individual model leads to a comparable predictive performance and a drastic reduction in computational costs. Further, the compressed individual model with best average rank is a rule-based regression model. Thus, model compression also leads to benefits in terms of model interpretability. The experiments carried in this paper are fully reproducible.

Cetin, B. and Yavuz, I. (2021). "Comparison of forecast accuracy of Ata and exponential smoothing." In: *Journal of Applied Statistics* 48(13-15), pp. 2580–2590.
Forecasting is a crucial step in almost all scientific research and is essential in many areas of industrial, commercial, clinical and economic activity. There are many forecasting methods in the literature; but exponential smoothing stands out due to its simplicity and accuracy. Despite the facts that exponential smoothing is widely used and has been in the literature for a long time, it suffers from some problems that potentially affect the model's forecast accuracy. An alternative forecasting framework, called Ata, was recently proposed to overcome these problems and to provide improved forecasts. In this study, the forecast accuracy of Ata and exponential smoothing will be compared among data sets with no or linear trend. The results of this study are obtained using simulated data sets with different sample sizes, variances. Forecast errors are compared within both short and long term forecasting horizons. The results show that the proposed approach outperforms exponential smoothing for both types of time series data when forecasting the near and distant future. The methods are implemented on the U.S. annualized monthly interest rates for services data and their forecasting performance are also compared for this data set.

Chai, D., Chiah, M., and Gharghori, P. (2019). "Which model best explains the returns of large Australian stocks?" In: *Pacific-Basin Finance Journal* 55, pp. 182–191.
Equity markets outside the US are generally dominated by small-sized stocks that are outside the investable universe of institutional investors and professional money managers. In this paper, we compare the performance of a range of competing factor models in pricing large Australian stocks. By doing so, we shed light on the mixed findings in prior studies and the issue of national and international pricing of assets. Using a comprehensive sample spanning a period of 35 years, we document that the Fama and French (2015) five-factor model is superior despite a few close matches with some of the competing models. As the sample expands from the top 300 to the top 500 stocks, the superiority of the five-factor model becomes more apparent. There is also evidence that profitability and investment factors help to explain the cross-section of stock returns. Finally, although large Australian stocks are integrated with the US market, domestic factors are more important drivers of expected returns in Australia.

Challu, C., Olivares, K. G., Welter, G., and Dubrawski, A. (2021). "DMIDAS: Deep Mixed Data Sampling Regression for Long Multi-Horizon Time Series Forecasting." In: *arXiv e-Print*.
Neural forecasting has shown significant improvements in the accuracy of large-scale systems, yet predicting extremely long horizons remains a challenging task. Two common problems are the volatility of the predictions and their computational complexity; we addressed them by incorporating smoothness regularization and mixed data sampling techniques to a well-performing multi-layer perceptron based architecture (NBEATS). We validate our proposed method, DMIDAS, on high-frequency healthcare and electricity price data with long forecasting horizons (approx 1000 timestamps) where we improve the prediction accuracy by 5% over state-of-the-art models, reducing the number of parameters of NBEATS by nearly 70%.

Chan, F. and Pauwels, L. L. (2018). "Some theoretical results on forecast combinations." In: *International Journal of Forecasting* 34(1), pp. 64–74.

This paper proposes a framework for the analysis of the theoretical properties of forecast combination, with the forecast performance being measured in terms of mean squared forecast errors (MSFE). Such a framework is useful for deriving all existing results with ease. In addition, it also provides insights into two forecast combination puzzles. Specifically, it investigates why a simple average of forecasts often outperforms forecasts from single models in terms of MSFEs, and why a more complicated weighting scheme does not always perform better than a simple average. In addition, this paper presents two new findings that are particularly relevant in practice. First, the MSFE of a forecast combination decreases as the number of models increases. Second, the conventional approach to the selection of optimal models, based on a simple comparison of MSFEs without further statistical testing, leads to a biased selection.

Charles, A., Darné, O., and Kim, J. H. (2022). "Stock return predictability: Evaluation based on interval forecasts." In: *Bulletin of Economic Research* 74(2), pp. 363–385.

This paper evaluates the predictability of monthly stock return using out-of-sample interval forecasts. Past studies exclusively use point forecasts, which are of limited value since they carry no information about intrinsic predictive uncertainty. We compare the empirical performance of alternative interval forecasts for stock return generated from a naive model, univariate autoregressive model, and multivariate model (predictive regression and VAR), using U.S. data from 1926. It is found that neither univariate nor multivariate interval forecasts outperform naive forecasts. This strongly suggests that the U.S. stock market has been informationally efficient in the weak form as well as in the semistrong form.

Charte, F., Vico, A., Perez-Godoy, M. D., and Rivera, A. J. (2019). "predtoolsTS: R package for streamlining time series forecasting." In: *Progress in Artificial Intelligence* 8(4), pp. 505–510.

Time series forecasting is a field of interest in many areas. Classically, statistical methods have been used to address this problem. In recent years, machine learning (ML) algorithms have been also applied with satisfactory results. However, ML software packages are not skilled to deal with raw sequences of temporal data, and therefore, it is necessary to transform these time series. This paper presents predtoolsTS, an R package that provides a uniform interface for applying both statistical and ML methods to time series forecasting. predtoolsTS comprises four modules: preprocessing, modeling, prediction and postprocessing, in order to deal with the whole process of time series forecasting.

Chatigny, P., Wang, S., Patenaude, J.-M., and Oreshkin, B. N. (2021). "Neural forecasting at scale." In: *arXiv e-Print*.

We study the problem of efficiently scaling ensemble-based deep neural networks for time series (TS) forecasting on a large set of time series. Current state-of-the-art deep ensemble models have high memory and computational requirements, hampering their use to forecast millions of TS in practical scenarios. We propose N-BEATS(P), a global multivariate variant of the N-BEATS model designed to allow simultaneous training of multiple univariate TS forecasting models. Our model addresses the practical limitations of related models, reducing the training time by half and memory requirement by a factor of 5, while keeping the same level of accuracy. We have performed multiple experiments detailing the various ways to train our model and have obtained results that demonstrate its capacity to support zero-shot TS forecasting, i.e., to train a neural network on a source TS dataset and deploy it on a different target TS dataset without retraining, which provides an efficient and reliable solution to forecast at scale even in difficult forecasting conditions.

Chatterjee, A., Bhowmick, H., and Sen, J. (2021). "Stock Price Prediction Using Time Series, Econometric, Machine Learning, and Deep Learning Models." In: *arXiv e-Print*.

For a long-time, researchers have been developing a reliable and accurate predictive model for stock price prediction. According to the literature, if predictive models are correctly designed and refined, they can painstakingly and faithfully estimate future stock values. This paper demonstrates a set of time series, econometric, and various learning-based models for stock price prediction. The data of Infosys, ICICI, and SUN PHARMA from the period of January 2004 to December 2019 was used here for training and testing the models to know which model performs best in which sector. One time series model (Holt-Winters Exponential Smoothing), one econometric model (ARIMA), two machine Learning models (Random Forest and MARS), and two deep learning-based models (simple RNN and LSTM) have been included in this paper. MARS has been proved to be the best performing machine learning model, while LSTM has proved to be the best performing deep learning model. But overall, for all three sectors - IT (on Infosys data), Banking (on ICICI data), and Health (on SUN PHARMA data), MARS has proved to be the best performing model in sales forecasting.

Chen, A. Y. and Zimmermann, T. (2021). "Open Source Cross-Sectional Asset Pricing." In: *SSRN e-Print*.

We provide data and code that successfully reproduces nearly all cross-sectional stock return predictors. Unlike most metastudies, we carefully examine the original papers to determine whether our predictability tests should produce t-stats above 1.96. For the 180 predictors that were clearly significant in the original papers, 98% of our reproductions find t-stats above 1.96. For the 30 predictors that had mixed evidence, our reproductions find t-stats of 2 on average. We include an additional 105 characteristics and 945 portfolios with alternative rebalancing frequencies to nest variables used in other metastudies. Our data covers all portfolios in Hou, Xue and Zhang (2017); 98% of the portfolios in McLean and Pontiff (2016); 90% of the characteristics from Green, Hand, and Zhang (2017); and 90% of the firm-level predictors in Harvey, Liu, and Zhu (2016) that use widely-available data.

Chen, H., Jiang, L., and Liu, W. (2020). "Predicting returns out of sample: A naive model averaging approach." In: *SSRN e-Print*.

The seminal paper by Goyal and Welch (2008) shows that variables that can forecast market returns in sample do not beat historical averages in forecasting market returns out of sample (i.e., the out-of-sample R2s are mostly negative). We reexamine this issue and present four findings: (i) A naive model averaging (NMA) method, by equally weighting the ordinary least squares (OLS) out-of-sample forecasts and the historical means, produces positive out-of-sample R2s for the variables that are significant in sample. (ii) The NMA method is helpful even after we impose additional restrictions, as in Campbell and Thompson (2008). When constructing composition forecasts based on multiple forecasting variables, our method of using historical means information is also useful relative to the method in Rapach, Strauss, and Zhou (2010). (iii) The Bayesian model averaging (BMA) approach fails to perform better than the NMA method. (iv) Declining return predictability does not explain the performance of the NMA method, which might be better explained by model misspecification.

Chen, L., Chen, D., Shang, Z., Zhang, Y., Wen, B., and Yang, C. (2022). "Multi-Scale Adaptive Graph Neural Network for Multivariate Time Series Forecasting." In: *arXiv e-Print*.

Multivariate time series (MTS) forecasting plays an important role in the automation and optimization of intelligent applications. It is a challenging task, as we need to consider both complex intra-variable dependencies and inter-variable dependencies. Existing works only learn temporal patterns with the help of single inter-variable dependencies. However, there are multi-scale temporal patterns in many real-world MTS. Single inter-variable dependencies make the model prefer to learn one type of prominent and shared temporal patterns. In this paper, we propose a multi-scale adaptive graph neural network (MAGNN) to address the above issue. MAGNN exploits a multi-scale pyramid network to preserve the underlying temporal dependencies at different time scales. Since the inter-variable dependencies may be different under distinct time scales, an adaptive graph learning module is designed to infer the scale-specific inter-variable dependencies without pre-defined priors. Given the multi-scale feature representations and scale-specific inter-variable dependencies, a multi-scale temporal graph neural network is introduced to jointly model intra-variable dependencies and inter-variable dependencies. After that, we develop a scale-wise fusion module to effectively promote the collaboration across different time scales, and automatically capture the importance of contributed temporal patterns. Experiments on four real-world datasets demonstrate that MAGNN outperforms the state-of-the-art methods across various settings.

Chen, L., Chen, W., Wu, B., Zhang, Y., Wen, B., and Yang, C. (2021a). "Learning from Multiple Time Series: A Deep Disentangled Approach to Diversified Time Series Forecasting." In: *arXiv e-Print*.

Time series forecasting is a significant problem in many applications, e.g., financial predictions and business optimization. Modern datasets can have multiple correlated time series, which are often generated with global (shared) regularities and local (specific) dynamics. In this paper, we seek to tackle such forecasting problems with DeepDGL, a deep forecasting model that disentangles dynamics into global and local temporal patterns. DeepDGL employs an encoder-decoder architecture, consisting of two encoders to learn global and local temporal patterns, respectively, and a decoder to make multi-step forecasting. Specifically, to model complicated global patterns, the vector quantization (VQ) module is introduced, allowing the global feature encoder to learn a shared codebook among all time series. To model diversified and heterogenous local patterns, an adaptive parameter generation module enhanced by the contrastive multi-horizon coding (CMC) is proposed to generate the parameters of the local feature encoder for each individual time series, which maximizes the mutual information between the series-specific context variable and the long/short-term representations of the corresponding time series. Our experiments on several real-world datasets show that DeepDGL outperforms existing state-of-the-art models.

Chen, L., Pelger, M., and Zhu, J. (2021b). "Deep learning in asset pricing." In: *SSRN e-Print*.

We use deep neural networks to estimate an asset pricing model for individual stock returns that takes advantage of the vast amount of conditioning information, while keeping a fully flexible form and accounting for time-variation. The key innovations are to use the fundamental no-arbitrage condition as criterion function, to construct the most informative test assets with an adversarial approach and to extract the states of the economy from many macroeconomic time series. Our asset pricing model outperforms out-of-sample all benchmark approaches in terms of Sharpe ratio, explained variation and pricing errors and identifies the key factors that drive asset prices. Code and data available at `https://mpelger.people.stanford.edu/data-and-code`.

Cheng, D., Yang, F., Xiang, S., and Liu, J. (2022). "Financial time series forecasting with multi-modality graph neural network." In: *Pattern Recognition* 121, p. 108218.
Financial time series analysis plays a central role in hedging market risks and optimizing investment decisions. This is a challenging task as the problems are always accompanied by multi-modality streams and lead-lag effects. For example, the price movements of stock are reflections of complicated market states in different diffusion speeds, including historical price series, media news, associated events, etc. Furthermore, the financial industry requires forecasting models to be interpretable and compliant. Therefore, in this paper, we propose a multi-modality graph neural network (MAGNN) to learn from these multimodal inputs for financial time series prediction. The heterogeneous graph network is constructed by the sources as nodes and relations in our financial knowledge graph as edges. To ensure the model interpretability, we leverage a two-phase attention mechanism for joint optimization, allowing end-users to investigate the importance of inner-modality and inter-modality sources. Extensive experiments on real-world datasets demonstrate the superior performance of MAGNN in financial market prediction. Our method provides investors with a profitable as well as interpretable option and enables them to make informed investment decisions.

Cheng, M., Swanson, N. R., and Yao, C. (2019). "Forecast Evaluation." In: *SSRN e-Print*.
The development of new tests and methods used in the evaluation of time series forecasts and forecasting models remains as important today as it has for the last 50 years. Paraphrasing what Sir Clive W.J. Granger (arguably the father of modern day time series forecasting) said in the 1990s at a conference in Svinkloev, Denmark, , the model looks like an interesting extension, but can it forecast better than existing models. Indeed, the forecast evaluation literature continues to expand, with interesting new tests and methods being developed at a rapid pace. In this chapter, we discuss a select a group of predictive accuracy tests and model selection methods that have been developed in recent years, and that are now widely used in the forecasting literature. We begin by reviewing several tests for comparing the relative forecast accuracy of different models, in the case of point forecasts. We then broaden the scope of our discussion by introducing density-based predictive accuracy tests. We conclude by noting that predictive accuracy is typically assessed in terms of a given loss function, such as mean squared forecast error or mean absolute forecast error. Most tests, including those discussed here, are consequently loss function dependent, and the relative forecast superiority of predictive models is therefore also dependent on specification of a loss function. In light of this fact, we conclude this chapter by discussing loss function robust predictive density accuracy tests that have recently been developed using principles of stochastic dominance.

Chevallier, J., Guégan, D., and Goutte, S. (2021). "Is It Possible to Forecast the Price of Bitcoin?" In: *Forecasting* 3(2), pp. 377–420.
This paper focuses on forecasting the price of Bitcoin, motivated by its market growth and the recent interest of market participants and academics. We deploy six machine learning algorithms (e.g., Artificial Neural Network, Support Vector Machine, Random Forest, k-Nearest Neighbours, AdaBoost, Ridge regression), without deciding a priori which one is the 'best' model. The main contribution is to use these data analytics techniques with great caution in the parameterization, instead of classical parametric modelings (AR), to disentangle the non-stationary behavior of the data. As soon as Bitcoin is also used for diversification in portfolios, we need to investigate its interactions with stocks, bonds, foreign exchange, and commodities. We identify that other cryptocurrencies convey enough information to explain the daily variation of Bitcoin's spot and futures prices. Forecasting results point to the segmentation of Bitcoin concerning alternative assets. Finally, trading strategies are implemented.

Chiah, M., Chai, D., Zhong, A., and Li, S. (2016). "A Better Model? An Empirical Investigation of the Fama-French Five-factor Model in Australia." In: *International Review of Finance* 16(4), pp. 595–638.
Recently, Fama and French (2015a) propose a five-factor model by adding profitability and investment factors to their three-factor model. This model outperforms the three-factor model previously proposed by Fama and French (1993). Using an extensive sample over the 1982-2013 period, we investigate the performance of the

five-factor model in pricing Australian equities. We find that the five-factor model is able to explain more asset pricing anomalies than a range of competing asset pricing models, which supports the superiority of the five-factor model. We also find that despite the results documented by Fama and French (2015a), the book-to-market factor retains its explanatory power in the presence of the investment and profitability factors. Our results are robust to alternative factor definitions and the formation of test assets. The study provides a strong out-of-sample test of the model, adding to the comparative evidence across international equity markets.

Chiang, I.-H. E., Liao, Y., and Zhou, Q. (2021). "Modeling the cross-section of stock returns using sensible models in a model pool." In: *Journal of Empirical Finance* 60, pp. 56–73.
An increase in the number of asset pricing models intensifies model uncertainties in asset pricing. While a pure ”model selection” (singling out a best model) can result in a loss of useful information, a full ”model pooling” may increase the risk of including noisy information. We make a trade-off between the two methods and develop a new two-step trimming-then-pooling method to forecast the joint distributions of asset returns using a large pool of asset pricing models. Our method allows investors to focus on certain regions of the distributions. In the first step, we trim the uninformative models from a pool of candidates, and in the second step, we pool the forecasts of the surviving models. We find that our method significantly enhances portfolio performance and predicts downside risk precisely, and the improvements are mainly due to trimming. The pool of sensible models becomes larger when focusing on extreme events, responds rapidly to rising uncertainty, and reflects the magnitude of factor premiums. These findings provide new insights into asset pricing model evaluation.

Chib, S. and Zeng, X. (2020). "Which factors are risk factors in asset pricing? A model scan framework." In: *Journal of Business & Economic Statistics* 38(4), pp. 771–783.
A key question for understanding the cross-section of expected returns of financial equity assets is the following: which factors, from a given collection of factors, are risk factors, equivalently, which factors are in the stochastic discount factor (SDF)? Though the SDF is unobserved, assumptions about which factors (from the available set of factors) are in the SDF restricts the joint distribution of factors in specific ways, as a consequence of the economic theory of asset pricing. A different starting collection of factors that go into the SDF leads to a different set of restrictions on the joint distribution of factors. The conditional distribution of equity returns has the same restricted form, regardless of what is assumed about the factors in the SDF, as long as the factors are tradeable, and hence the distribution of asset returns is irrelevant for isolating the risk-factors. The restricted factors models are distinct (non-nested) and do not arise by omitting or including a variable from a full model, thus precluding analysis by standard statistical variable selection methods, such as those based on the lasso and its variants. Instead, we develop what we call a Bayesian model scan strategy in which each factor is allowed to enter or not enter the SDF and the resulting restricted models (of which there are 114,674 in our empirical study) are simultaneously confronted with the data. We use a student-t distribution for the factors, and model-specific independent student-t distribution for the location parameters, a training sample to fix prior locations, and a creative way to arrive at the joint distribution of several other model-specific parameters from a single prior distribution. This allows our method to be essentially a scaleable and tuned-black-box method that can be applied across our large model space with little to no user-intervention. The model marginal likelihoods, and implied posterior model probabilities, are compared with the prior probability of 1/114,674 of each model to find the best supported model, and thus the factors most likely to be in the SDF. We provide detailed simulation evidence about the high finite-sample accuracy of the method. Our empirical study with 13 leading factors reveals that the highest marginal likelihood model is a student-t distributed factor model with 5 degrees of freedom and 8 risk factors.

Chib, S., Zeng, X., and Zhao, L. (2020). "On Comparing Asset Pricing Models." In: *Journal of Finance* 75(11), pp. 551–577.
We revisit the framework of Barillas and Shanken (2018) (BS henceforth) to point out that the Bayesian marginal likelihood based model comparison method in that paper is unsound. We show that in this comparison of asset pricing models, the priors on the nuisance parameters across models must satisfy a certain change of variable property for densities that is violated by the off-the-shelf Jeffreys priors used in the BS method. Hence, the BS ”marginal likelihoods” are non-comparable across models and cannot be used to locate the (traded) risk factors. We conduct extensive simulation exercises in two designs: one with 8 potential pricing factors and a second with 12 factors, in each case matching the factors to real world factors that arise in this setting. As expected, the BS method performs unsatisfactorily, even when epic (and practically unattainable) sample sizes of .12 and 1.2 million are used to conduct the model comparisons. In a notable advance, we derive a new class of improper priors on the nuisance parameters, starting from a single improper prior, which leads to valid marginal

likelihoods, and valid model comparisons. The empirical performance of our marginal likelihoods is substantially better, opening doors to reliable Bayesian work on which factors are risk factors in asset pricing models.

Chib, S., Zhao, L., Huang, D., and Zhou, G. (2022). "Winners from Winners: A Tale of Risk Factors." In: *SSRN e-Print*.

Taking the union of the risk factors recently proposed by Fama and French (1993, 2015, 2018), Hou, Xue, and Zhang (2015), Stambaugh and Yuan (2017), and Daniel, Hirshleifer, and Sun (2019), a pool we refer to as the , we ask what collection of winners from winners emerge when each factor is allowed to play the role of a risk factor, or a non-risk factor. Our comparison of 4,095 models shows that a six factor model consisting of Mkt, SMB, MOM, ROE, MGMT, and PEAD as risk factors has the largest Bayesian posterior probability. Moreover, this collection displays superior out-of-sample predictive performance, higher Sharpe ratios, and greater ability in pricing anomalies, than the preceding models. These results suggest that both fundamental and behavioral factors play an important role in explaining the cross-section of expected equity returns.

Chiu, C.-W., Hayes, S., Kapetanios, G., and Theodoridis, K. (2019). "A new approach for detecting shifts in forecast accuracy." In: *International Journal of Forecasting* 35(4), pp. 1596–1612.

Forecasts play a critical role at inflation-targeting central banks, such as the Bank of England. Breaks in the forecast performance of a model can potentially incur important policy costs. Commonly used statistical procedures, however, implicitly put a lot of weight on type I errors (or false positives), which result in a relatively low power of tests to identify forecast breakdowns in small samples. We develop a procedure which aims at capturing the policy cost of missing a break. We use data-based rules to find the test size that optimally trades off the costs associated with false positives with those that can result from a break going undetected for too long. In so doing, we also explicitly study forecast errors as a multivariate system. The covariance between forecast errors for different series, though often overlooked in the forecasting literature, not only enables us to consider testing in a multivariate setting but also increases the test power. As a result, we can tailor the choice of the critical values for each series not only to the in-sample properties of each series but also to how the series for forecast errors covary.

Choe, Y. J. and Ramdas, A. (2022). "Comparing Sequential Forecasters." In: *arXiv e-Print*.

Consider two or more forecasters, each making a sequence of predictions for different events over time. We ask a relatively basic question: how might we compare these forecasters, either online or post-hoc, while avoiding unverifiable assumptions on how the forecasts or outcomes were generated? This work presents a novel and rigorous answer to this question. We design a sequential inference procedure for estimating the time-varying difference in forecast quality as measured by any scoring rule. The resulting confidence intervals are nonasymptotically valid and can be continuously monitored to yield statistically valid comparisons at arbitrary data-dependent stopping times ("anytime-valid"); this is enabled by adapting variance-adaptive supermartingales, confidence sequences, and e-processes to our setting. Motivated by Shafer and Vovk's game-theoretic probability, our coverage guarantees are also distribution-free, in the sense that they make no distributional assumptions on the forecasts or outcomes. In contrast to a recent work by Henzi and Ziegel, our tools can sequentially test a weak null hypothesis about whether one forecaster outperforms another on average over time. We demonstrate their effectiveness by comparing probability forecasts on Major League Baseball (MLB) games and statistical postprocessing methods for ensemble weather forecasts.

Cholakov, R. and Kolev, T. (2021). "Transformers predicting the future. Applying attention in next-frame and time series forecasting." In: *arXiv e-Print*.

Recurrent Neural Networks were, until recently, one of the best ways to capture the timely dependencies in sequences. However, with the introduction of the Transformer, it has been proven that an architecture with only attention-mechanisms without any RNN can improve on the results in various sequence processing tasks (e.g. NLP). Multiple studies since then have shown that similar approaches can be applied for images, point clouds, video, audio or time series forecasting. Furthermore, solutions such as the Perceiver or the Informer have been introduced to expand on the applicability of the Transformer. Our main objective is testing and evaluating the effectiveness of applying Transformer-like models on time series data, tackling susceptibility to anomalies, context awareness and space complexity by fine-tuning the hyperparameters, preprocessing the data, applying dimensionality reduction or convolutional encodings, etc. We are also looking at the problem of next-frame prediction and exploring ways to modify existing solutions in order to achieve higher performance and learn generalized knowledge.

Chordia, T., Goyal, A., and Saretto, A. (2020). "Anomalies and false rejections." In: *The Review of Financial Studies* 33(5), pp. 2134–2179.

We use information from over 2 million trading strategies randomly generated using real data and from strategies that survive the publication process to infer the statistical properties of the set of strategies that could have been studied by researchers. Using this set, we compute t-statistic thresholds that control for multiple hypothesis testing, when searching for anomalies, at 3.8 and 3.4 for time-series and cross-sectional regressions, respectively. We estimate the expected proportion of false rejections that researchers would produce if they failed to account for multiple hypothesis testing to be about 45%.

Chu, P. K. (2021). "Forecasting Recessions with Financial Variables and Temporal Dependence." In: *Economies* 9(3), p. 118.
Extending earlier research on forecasting recessions with financial variables, I examine the importance of additional financial variables and temporal dependence for recession prediction. I show that both additional financial variables, in particular, the Treasury bill spread, default yield spread, stock return volatility, and temporal cubic terms, which account for temporal dependence, independently help to improve not only in-sample, but also out-of-sample recession prediction. I also find that additional financial variables and temporal cubic terms complement each other in enhancing the predictability of recessions, increasing the explanatory power and decreasing prediction error further, compared to their individual performance.

Chudik, A., Pesaran, M. H., and Sharifvaghefi, M. (2021). "Variable Selection and Forecasting in High Dimensional Linear Regressions with Structural Breaks." In: *SSRN e-Print* 2020(394).
This paper is concerned with the problem of variable selection and forecasting in the presence of parameter instability. There are a number of approaches proposed for forecasting in the presence of breaks, including the use of rolling windows and exponential down-weighting. However, these studies start with a given model specification and do not consider the problem of variable selection, which is complicated by time variations in the effects of signal variables. In this study we investigate whether or not we should use weighted observations at the variable selection stage in the presence of structural breaks, particularly when the number of potential covariates is large. Amongst the extant variable selection approaches we focus on the recently developed One Covariate at a time Multiple Testing (OCMT) method. This procedure allows a natural distinction between the selection and forecasting stages. We establish three main theorems on selection, estimation post selection and in-sample fit. These theorems provide justification for using the full (not down-weighted) sample at the selection stage of OCMT and down-weighting of observations only at the forecasting stage (if needed). The benefits of the proposed method are illustrated by empirical applications to forecasting output growths and stock market returns.

Cohen, N., Sood, S., Zeng, Z., Balch, T., and Veloso, M. (2020). "Visual Forecasting of Time Series with Image-to-Image Regression." In: *arXiv e-Print*.
Time series forecasting is essential for agents to make decisions in many domains. Existing models rely on classical statistical methods to predict future values based on previously observed numerical information. Yet, practitioners often rely on visualizations such as charts and plots to reason about their predictions. Inspired by the end-users, we re-imagine the topic by creating a framework to produce visual forecasts, similar to the way humans intuitively do. In this work, we take a novel approach by leveraging advances in deep learning to extend the field of time series forecasting to a visual setting. We do this by transforming the numerical analysis problem into the computer vision domain. Using visualizations of time series data as input, we train a convolutional autoencoder to produce corresponding visual forecasts. We examine various synthetic and real datasets with diverse degrees of complexity. Our experiments show that visual forecasting is effective for cyclic data but somewhat less for irregular data such as stock price. Importantly, we find the proposed visual forecasting method to outperform numerical baselines. We attribute the success of the visual forecasting approach to the fact that we convert the continuous numerical regression problem into a discrete domain with quantization of the continuous target signal into pixel space.

Cohen, N., Sood, S., Zeng, Z., Balch, T., and Veloso, M. (2021). "Visual Time Series Forecasting: An Image-driven Approach." In: *MiLeTS'21: 7th KDD Workshop on Mining and Learning from Time Series*.
In this work, we address time-series forecasting as a computer vision task. We capture input data as an image and train a model to produce the subsequent image. This approach results in predicting distributions as opposed to pointwise values. To assess the robustness and quality of our approach, we examine various datasets and multiple evaluation metrics. Our experiments show that our forecasting tool is effective for cyclic data but somewhat less for irregular data such as stock prices. Importantly, when using image-based evaluation metrics, we find our method to outperform various baselines, including ARIMA, and a numerical variation of our deep learning approach.

Collot, S. and Hemauer, T. (2021). "A literature review of new methods in empirical asset pricing: omitted-variable and errors-in-variable bias." In: *Financial Markets and Portfolio Management* 35, pp. 77–100.
Standard procedures in empirical asset pricing suffer from various issues that are common to all regression-based methods. This work reviews recently introduced approaches that aim to mitigate problems associated with omitted factors and errors-in-variables. New methods addressing the omitted-variable bias suggest procedures for selecting appropriate control variables, aggregating the information from a large set of factors, or making existing methods robust against omitted factors. While the omitted-variable problem is present in almost all standard empirical asset pricing methods, the errors-in-variables problem is largely limited to the estimation of factor premia via two-pass regressions. New methods addressing the errors-in-variable bias implement an instrumental variable approach, suggest a generalized version of the widely used portfolio sorts procedure, or correct estimates based on an analytic expression for the bias. Ultimately, all of these new methods represent highly relevant advances for the area of empirical asset pricing, and the possibility to synthesize the most promising approaches might be worthwhile to investigate in the future.

Cong, L. W., Tang, K., Wang, J., and Zhang, Y. (2021). "Deep Sequence Modeling: Development and Applications in Asset Pricing." In: *The Journal of Financial Data Science* 3(1), pp. 28–42.
The authors predict asset returns and measure risk premiums using a prominent technique from artificial intelligence: deep sequence modeling. Because asset returns often exhibit sequential dependence that may not be effectively captured by conventional time-series models, sequence modeling offers a promising path with its data-driven approach and superior performance. In this article the authors first overview the development of deep sequence models, introduce their applications in asset pricing, and discuss their advantages and limitations. They then perform a comparative analysis of these methods using data on U.S. equities. They demonstrate how sequence modeling benefits investors in general through incorporating complex historical path dependence and that long short-term memory-based models tend to have the best out-of-sample performance.

Coqueret, G. and Guida, T. (2020). *Machine Learning for Factor Investing: R Version*. Chapman and Hall/CRC. 341 pp.
The book covers a wide array of subjects which range from economic rationales to rigorous portfolio back-testing and encompass both data processing and model interpretability. Common supervised learning algorithms such as tree models and neural networks are explained in the context of style investing and the reader can also dig into more complex techniques like autoencoder asset returns, Bayesian additive trees, and causal models. All topics are illustrated with self-contained R code samples and snippets that are applied to a large public dataset that contains over 90 predictors. The material, along with the content of the book, is available online so that readers can reproduce and enhance the examples at their convenience. If you have even a basic knowledge of quantitative finance, this combination of theoretical concepts and practical illustrations will help you learn quickly and deepen your financial and technical expertise.

Cornell, B. (2020). "Stock characteristics and stock returns: a skeptic's look at the cross section of expected returns." In: *The Journal of Portfolio Management* 46(5), pp. 131–142.
The correlation between stock characteristics and the cross section of stock returns plays a central role in empirical implementations of modern asset pricing models and has important implications for investment management. This remains true whether the correlation is due to investor preferences regarding the characteristics directly or whether the characteristics are proxies for state variables, the risk of which investors are attempting to hedge. This article asks what we know about the relation between these characteristics and the cross section of returns. The skeptic answer is, not much. A combination of lack of persistence in the characteristics and problems caused by model uncertainty, data snooping, and nonstationarity means that our knowledge is sketchy at best. Investors should be forewarned when considering any strategies such as smart beta that are premised on the correlation between characteristics and the cross section of returns.

Coroneo, L., Iacone, F., Paccagnini, A., and Monteiro, P. S. (2021). "Testing the Predictive Accuracy of COVID-19 Forecasts." In: *SSRN e-Print*.
We test the predictive accuracy of forecasts of the number of COVID-19 fatalities produced by several forecasting teams and collected by the United States Centers for Disease Control and Prevention during the first and second waves of the epidemic in the United States. We find three main results. First, at the short horizon (1-week ahead) no forecasting team outperforms a simple time-series benchmark. Second, at longer horizons (3- and 4-week ahead) forecasters are more successful and sometimes outperform the benchmark, in particular during the first wave of the epidemic. Third, one of the best performing forecasts is the Ensemble forecast, that combines all available predictions using uniform weights. In view of these results, collecting a wide range of forecasts and

combining them in an ensemble forecast may be a superior approach for health authorities, rather than relying on a small number of forecasts.

Costantini, M. and Kunst, R. M. (2021). "On using predictive-ability tests in the selection of time-series prediction models: A Monte Carlo evaluation." In: *International Journal of Forecasting* 37(2), pp. 445–460.

To select a forecast model among competing models, researchers often use ex-ante prediction experiments over training samples. Following Diebold and Mariano (1995), forecasters routinely evaluate the relative performance of competing models with accuracy tests and may base their selection on test significance on top of comparing forecast errors. With extensive Monte Carlo analysis, we investigated whether this practice favors simpler models over more complex ones, without gains in forecast accuracy. We simulated the autoregressive moving-average model, the self-exciting threshold autoregressive model, and vector autoregression. We considered two variants of the Diebold-Mariano test, the test by Giacomini and White (2006), the -test by Clark and McCracken (2001), the Akaike information criterion, and a pure training-sample evaluation. The findings showed some accuracy gains for small samples when applying accuracy tests, particularly for the Clark-McCracken and bootstrapped Diebold-Mariano tests. Evidence against this testing procedure dominated, however, and training-sample evaluations without accuracy tests performed best in many cases.

Czasonis, M., Kritzman, M., and Turkington, D. (2020). "Addition by Subtraction: A Better Way to Forecast Factor Returns (and Everything Else)." In: *The Journal of Portfolio Management* 46(8), pp. 98–107.

Financial analysts assume that the reliability of predictions derived from regression analysis improves with sample size. This is thought to be true because larger samples tend to produce less noisy results than smaller samples. But this is not always the case. Some observations are more relevant than others, and often one can obtain more reliable predictions by censoring observations that are not sufficiently relevant. The authors introduce a methodology for identifying relevant observations by recasting the prediction of a regression equation as a weighted average of the historical values of the dependent variable, in which the weights are the relevance of the independent variables. This equivalence allows them to use a subset of more relevant observations to forecast the dependent variable. The authors apply their methodology to forecast factor returns from economic variables.

Czasonis, M., Kritzman, M., and Turkington, D. (2021a). "Relevance." In: *SSRN e-Print*.

The authors describe a new statistical concept called relevance from a conceptual and mathematical perspective, and based on their mathematical framework, they present a unified theory of relevance, regressions, and event studies. They also include numerical examples of how relevance is used to forecast.

Czasonis, M., Kritzman, M., and Turkington, D. (2021b). "The Past as Prologue: How to Forecast Presidential Elections." In: *SSRN e-Print*.

It is common practice to forecast social, political, and economic outcomes by polling people about their intentions. This approach is direct, but it can be unreliable in settings where it is hard to identify a representative sample, or where subjects have an incentive to conceal their true intentions or beliefs. The authors propose that, as a substitute or a supplement, forecasters use historical outcomes to predict future ones. The relevance of historical events, however, is not guaranteed. The authors apply a novel technique called Partial Sample Regression to identify, in a mathematically precise way, the subset of events that are most relevant to the present. The outcomes of those events are then weighted by their relevance and averaged to give a prediction for the future. The authors illustrate their technique by showing that it correctly predicted the winner of the last six U.S. presidential elections based only on the political, geopolitical, and economic circumstances of the election year.

Dai, Z., Kang, J., and Wen, F. (2021). "Predicting stock returns: A risk measurement perspective." In: *International Review of Financial Analysis* 74, p. 101676.

This paper proposes a new and efficient model selection strategy to obtain significant stock returns predictability from a risk measurement perspective. The risk interval is defined as the distance between the current actual return and the returns' historical average. The model selection strategy involves switching stock return forecasting models according to different risk intervals from the mean reversion and extreme value theory. This new strategy generates encouraging results in the empirical analysis. A mean-variance investor can realize sizeable economic gains by allocating assets through this new approach relative to competing forecasting models. Furthermore, the strategy performs robustly under alternative settings from both statistical and economic perspectives.

Dai, Z., Li, T., and Yang, M. (2022). "Forecasting stock return volatility: The role of shrinkage approaches in a data-rich environment." In: *Journal of Forecasting*.

This paper employs the prevailing shrinkage approaches, the lasso, adaptive lasso, elastic net and ridge regression to predict stock return volatility with a large set of variables. The out-of-sample results reveal that shrinkage approaches exhibit superior performance relative to the benchmark of the autoregressive model and a series of competing models in terms of the out-of-sample R-square and the model confidence set. By using shrinkage methods to allocate portfolio, a mean-variance investor can obtain significant economic gains. Overall, our findings confirm that shrinkage approaches can effectively improve stock return volatility forecasting in a data-rich environment.

Dama, F. and Sinoquet, C. (2021). "Analysis and modeling to forecast in time series: a systematic review." In: *arXiv e-Print*.
This paper surveys state-of-the-art methods and models dedicated to time series analysis and modeling, with the final aim of prediction. This review aims to offer a structured and comprehensive view of the full process flow, and encompasses time series decomposition, stationary tests, modeling and forecasting. Besides, to meet didactic purposes, a unified presentation has been adopted throughout this survey, to present decomposition frameworks on the one hand and linear and nonlinear time series models on the other hand. First, we decrypt the relationships between stationarity and linearity, and further examine the main classes of methods used to test for weak stationarity. Next, the main frameworks for time series decomposition are presented in a unified way: depending on the time series, a more or less complex decomposition scheme seeks to obtain nonstationary effects (the deterministic components) and a remaining stochastic component. An appropriate modeling of the latter is a critical step to guarantee prediction accuracy. We then present three popular linear models, together with two more flexible variants of the latter. A step further in model complexity, and still in a unified way, we present five major nonlinear models used for time series. Amongst nonlinear models, artificial neural networks hold a place apart as deep learning has recently gained considerable attention. A whole section is therefore dedicated to time series forecasting relying on deep learning approaches. A final section provides a list of R and Python implementations for the methods, models and tests presented throughout this review. In this document, our intention is to bring sufficient in-depth knowledge, while covering a broad range of models and forecasting methods: this compilation spans from well-established conventional approaches to more recent adaptations of deep learning to time series forecasting.

Davydenko, A. and Goodwin, P. (2021). "Assessing Point Forecast Bias Across Multiple Time Series: Measures and Visual Tools." In: *International Journal of Statistics and Probability* 10(5), p. 46.
Measuring bias is important as it helps identify flaws in quantitative forecasting methods or judgmental forecasts. It can, therefore, potentially help improve forecasts. Despite this, bias tends to be under represented in the literature: many studies focus solely on measuring accuracy. Methods for assessing bias in single series are relatively well known and well researched, but for datasets containing thousands of observations for multiple series, the methodology for measuring and reporting bias is less obvious. We compare alternative approaches against a number of criteria when rolling origin point forecasts are available for different forecasting methods and for multiple horizons over multiple series. We focus on relatively simple, yet interpretable and easy to implement metrics and visualization tools that are likely to be applicable in practice. To study the statistical properties of alternative measures we use theoretical concepts and simulation experiments based on artificial data with predetermined features. We describe the difference between mean and median bias, describe the connection between metrics for accuracy and bias, provide suitable bias measures depending on the loss function used to optimise forecasts, and suggest which measures for accuracy should be used to accompany bias indicators. We propose several new measures and provide our recommendations on how to evaluate forecast bias across multiple series.

De Baets, S. and Harvey, N. (2020). "Using judgment to select and adjust forecasts from statistical models." In: *European Journal of Operational Research* 284(3), pp. 882–895.
Forecasting support systems allow users to choose different statistical forecasting methods. But how well do they make this choice? We examine this in two experiments. In the first one (N=191), people selected the model that they judged to perform the best. Their choice outperformed forecasts made by averaging the model outputs and improved with a larger difference in quality between models and a lower level of noise in the data series. In a second experiment (N=161), participants were asked to make a forecast and were then offered advice in the form of a model forecast. They could then re-adjust their forecast. Final forecasts were more influenced by models that made better forecasts. As forecasters gained experience, they followed input from high-quality models more readily. Thus, both experiments show that forecasters have ability to use and learn from visual records of past performance to select and adjust model-based forecasts appropriately.

de Valk, S., Mattos, D. de, and Ferreira, P. (2019). "Nowcasting: An R Package for Predicting Economic Variables Using Dynamic Factor Models." In: *The R Journal*.

The nowcasting package provides the tools to make forecasts of monthly or quarterly economic variables using dynamic factor models. The objective is to help the user at each step of the forecasting process, starting with the construction of a database, all the way to the interpretation of the forecasts. The dynamic factor model adopted in this package is based on the articles from Giannone et al. (2008) and Banbura et al. (2011). Although there exist several other dynamic factor model packages available for R, ours provides an environment to easily forecast economic variables and interpret results.

Debnath, A., Waghmare, G., Wadhwa, H., Asthana, S., and Arora, A. (2021). "Exploring Generative Data Augmentation in Multivariate Time Series Forecasting : Opportunities and Challenges." In: *MiLeTS'21: 7th KDD Workshop on Mining and Learning from Time Series*.

In multivariate time series (MTS), each time point constitutes multiple time-dependent variables. Short-term and long-term correlation of these variables is a significant characteristic of MTS, and is a key challenge while modelling the same. While classical auto-regressive models are heavily used to model MTS, neural models are more flexible and efficient. However, neural models rely on a large amount of labelled data for training. Availability of labelled time series data could be a bottleneck in real-world scenarios. This scarcity of labelled data can be mitigated by data augmentation. In MTS, augmentation techniques need to realize short-term correlations and long-term temporal dynamics. In this work, we introduce a novel meta-algorithm for time-series data augmentation to address the data scarcity problem. Due to the intrinsic ordering of samples in time series, we argue that one cannot simply add synthetic samples to the real samples for augmentation. To this end, we generate synthetic MTS data preserving temporal dynamics using an offthe-shelf generative algorithm and frame augmentation in MTS as a transfer learning problem. In addition, we point out the drawbacks of generative model in MTS augmentation. We show the effectiveness of our method on publicly available MTS datasets in forecasting. We also perform qualitative and quantitative analysis of synthetic MTS data and its applicability in long-term forecasting. To the best of our knowledge, this is the first study on generative data augmentation for MTS forecasting.

Dendramis, Y., Kapetanios, G., and Marcellino, M. (2020). "A similarity-based approach for macroeconomic forecasting." In: *Journal of the Royal Statistical Society Series A* 183(3), pp. 801–827.

In the aftermath of the recent financial crisis there has been considerable focus on methods for predicting macroeconomic variables when their behaviour is subject to abrupt changes, associated for example with crisis periods. We propose similarity-based approaches as a way to handle parameter instability and apply them to macroeconomic forecasting. The rationale is that clusters of past data that match the current economic conditions can be more informative for forecasting than the entire past behaviour of the variable of interest. We apply our methods to predict both simulated data in a set of Monte Carlo experiments, and a broad set of key US macroeconomic indicators. The forecast evaluation exercises indicate that similarity-based approaches perform well, in general, in comparison with other common time-varying forecasting methods, and particularly well during crisis episodes.

Deshpande, P., Marathe, K., De, A., and Sarawagi, S. (2021). "Long Horizon Forecasting With Temporal Point Processes." In: *arXiv e-Print*.

In recent years, marked temporal point processes (MTPPs) have emerged as a powerful modeling machinery to characterize asynchronous events in a wide variety of applications. MTPPs have demonstrated significant potential in predicting event-timings, especially for events arriving in near future. However, due to current design choices, MTPPs often show poor predictive performance at forecasting event arrivals in distant future. To ameliorate this limitation, in this paper, we design DualTPP which is specifically well-suited to long horizon event forecasting. DualTPP has two components. The first component is an intensity free MTPP model, which captures microscopic or granular level signals of the event dynamics by modeling the time of future events. The second component takes a different dual perspective of modeling aggregated counts of events in a given time-window, thus encapsulating macroscopic event dynamics. Then we develop a novel inference framework jointly over the two models percentages for efficiently forecasting long horizon events by solving a sequence of constrained quadratic optimization problems. Experiments with a diverse set of real datasets show that DualTPP outperforms existing MTPP methods on long horizon forecasting by substantial margins, achieving almost an order of magnitude reduction in Wasserstein distance between actual events and forecasts.

Deshpande, P. and Sarawagi, S. (2021). "Long Range Probabilistic Forecasting in Time-Series using High Order Statistics." In: *arXiv e-Print*.

Long range forecasts are the starting point of many decision support systems that need to draw inference from high-level aggregate patterns on forecasted values. State of the art time-series forecasting methods are

either subject to concept drift on long-horizon forecasts, or fail to accurately predict coherent and accurate high-level aggregates. In this work, we present a novel probabilistic forecasting method that produces forecasts that are coherent in terms of base level and predicted aggregate statistics. We achieve the coherency between predicted base-level and aggregate statistics using a novel inference method. Our inference method is based on KL-divergence and can be solved efficiently in closed form. We show that our method improves forecast performance across both base level and unseen aggregates post inference on real datasets ranging three diverse domains.

Di Fonzo, T. and Girolimetto, D. (2020). "Cross-temporal forecast reconciliation: Optimal combination method and heuristic alternatives." In: *arXiv e-Print*.

Forecast reconciliation is a post-forecasting process aimed to improve the quality of the base forecasts for a system of hierarchical/grouped time series (Hyndman et al., 2011). Contemporaneous (cross-sectional) and temporal hierarchies have been considered in the literature, but - except for Kourentzes and Athanasopoulos (2019) - generally these two features have not been fully considered together. Adopting a notation able to simultaneously deal with both forecast reconciliation dimensions, the paper shows two new results: (i) an iterative cross-temporal forecast reconciliation procedure which extends, and overcomes some weaknesses of, the two-step procedure by Kourentzes and Athanasopoulos (2019), and (ii) the closed-form expression of the optimal (in least squares sense) point forecasts which fulfill both contemporaneous and temporal constraints. The feasibility of the proposed procedures, along with first evaluations of their performance as compared to the most performing 'single dimension' (either cross-sectional or temporal) forecast reconciliation procedures, is studied through a forecasting experiment on the 95 quarterly time series of the Australian GDP from Income and Expenditure sides considered by Athanasopoulos et al. (2019).

Di Fonzo, T. and Girolimetto, D. (2021). "Forecast combination based forecast reconciliation: insights and extensions." In: *arXiv e-Print*.

In a recent paper, while elucidating the links between forecast combination and cross-sectional forecast reconciliation, Hollyman et al. (2021) have proposed a forecast combination-based approach to the reconciliation of a simple hierarchy. A new Level Conditional Coherent (LCC) point forecast reconciliation procedure was developed, and it was shown that the simple average of a set of LCC, and bottom-up reconciled forecasts (called Combined Conditional Coherent, CCC) results in good performance as compared to those obtained through the state-of-the-art cross-sectional reconciliation procedures. In this paper, we build upon and extend this proposal along some new directions. (1) We shed light on the nature and the mathematical derivation of the LCC reconciliation formula, showing that it is the result of an exogenously linearly constrained minimization of a quadratic loss function in the differences between the target and the base forecasts with a diagonal associated matrix. (2) Endogenous constraints may be considered as well, resulting in level conditional reconciled forecasts of all the involved series, where both the upper and the bottom time series are coherently revised. (3) As the LCC procedure does not guarantee the non-negativity of the reconciled forecasts, we argue that - when non-negativity is a natural attribute of the variables to be forecast - its interpretation as an unbiased top-down reconciliation procedure leaves room for some doubts. (4) The new procedures are used in a forecasting experiment on the classical Australian Tourism Demand (Visitor Nights) dataset. Due to the crucial role played by the (possibly different) models used to compute the base forecasts, we re-interpret the CCC reconciliation of Hollyman et al. (2021) as a forecast pooling approach, showing that accuracy improvement may be gained by adopting a simple forecast averaging strategy.

Di Fonzo, T. and Girolimetto, D. (2022). "Cross-temporal forecast reconciliation: Optimal combination method and heuristic alternatives." In: *International Journal of Forecasting*.

Forecast reconciliation is a post-forecasting process aimed to improve the quality of the base forecasts for a system of hierarchical/grouped time series. Cross-sectional and temporal hierarchies have been considered in the literature, but generally, these two features have not been fully considered together. The paper presents two new results by adopting a notation that simultaneously deals with both forecast reconciliation dimensions. (i) The closed-form expression of the optimal (in the least squares sense) point forecasts fulfilling both contemporaneous and temporal constraints. (ii) An iterative procedure that produces cross-temporally reconciled forecasts by alternating forecast reconciliation along one single dimension (either cross-sectional or temporal) at each iteration step. The feasibility of the proposed procedures, along with first evaluations of their performance as compared to the most performing 'single dimension' (either cross-sectional or temporal) forecast reconciliation procedures, is studied through a forecasting experiment on the 95 quarterly time series of the Australian Gross Domestic Product from Income and Expenditure sides. For this dataset, the new procedures, in addition to providing

fully coherent forecasts in both cross-sectional and temporal dimensions, improve the forecast accuracy of the state-of-the-art point forecast reconciliation techniques. Implementation in R Package FoReCo on CRAN at https://cran.r-project.org/web/packages/FoReco/index.html.

Diebold, F. X. (2015). "Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold-Mariano Tests." In: *Journal of Business and Economic Statistics* 33(1), p. 1.
The Diebold-Mariano (DM) test was intended for comparing forecasts; it has been, and remains, useful in that regard. The DM test was not intended for comparing models. Much of the large ensuing literature, however, uses DM-type tests for comparing models, in pseudo-out-of-sample environments. In that case, simpler yet more compelling full-sample model comparison procedures exist; they have been, and should continue to be, widely used. The hunch that pseudo-out-of-sample analysis is somehow the only , or best , or even necessarily a good way to provide insurance against in-sample overfitting in model comparisons proves largely false. On the other hand, pseudo-out-of-sample analysis remains useful for certain tasks, perhaps most notably for providing information about comparative predictive performance during particular historical episodes.

Dong, H., Guo, X., Reichgelt, H., and Hu, R. (2020). "Predictive power of ARIMA models in forecasting equity returns: a sliding window method." In: *Journal of Asset Management* (21), pp. 549–566.
The ARIMA model is widely adopted by the financial industry as the standard statistical instrument for forecasting asset returns. Numerous studies have compared the accuracy of the ARIMA model with other competing models. However, there are no studies that cover a broad range of equities and their time series. Furthermore, there is no clear guideline on the time series window selected to fit the ARIMA model. In addition, there are no firm conclusions on whether older information in the sample should be abandoned. This makes it impossible to draw a definitive conclusion about the predictive power of the ARIMA model. This study sets out to address this gap in the literature. It summarizes more than two million ARIMA forecasts of future daily returns, using data from January 3, 1996 to May 12, 2017. The forecasts are run with different model parameter settings. We find that the five-year sliding fixed-width window fits US equity market asset prices to the highest degree, with an annual over-optimistic error of 2.6561%. However, when environments with positive and negative returns are separated, the ARIMA models generate forecasting errors of -0.0009% and 0.011%, and both underestimate gain and loss. These errors are lower for low volatility equities. We conclude that the lack of nonlinearity of the ARIMA model is not a major concern, and that the ARIMA models do not lose their validity if the data windows are carefully selected. Our conclusions are not in conflict with the weak form market efficiency hypothesis and are robust in an environment with transaction cost.

Dong, X., Li, Y., Rapach, D., and Zhou, G. (2022). "Anomalies and the expected market return." In: *Journal of Finance* 27(1), pp. 639–681.
We provide the first systematic evidence on the link between long-short anomaly portfolio returns (a cornerstone of the cross-sectional literature) and the time-series predictability of the aggregate market excess return. Using 100 representative anomalies from the literature, we employ a variety of shrinkage techniques (including machine learning, forecast combination, and dimension reduction) to efficiently extract predictive signals in a high-dimensional setting. We find that long-short anomaly portfolio returns evince statistically and economically significant out-of-sample predictive ability for the market excess return. Economically, the predictive ability of anomaly portfolio returns appears to stem from asymmetric limits of arbitrage and overpricing dominance.
Code and data is provided as supplement to article.

Drobetz, W. and Otto, T. (2020). "Empirical Asset Pricing via Machine Learning: Evidence from the European Stock Market." In: *SSRN e-Print*.
This paper evaluates the performance of machine learning methods in forecasting stock returns. Compared to a linear benchmark model, interactions and non-linear effects help improve predictive performance. But machine learning models must be adequately trained and tuned to overcome the high dimensionality issue and to avoid over-fitting. Across all machine learning methods, the most important predictors are based on price trends and fundamental signals from valuation ratios. However, the models exhibit disparities in statistical performance that translate into pronounced differences in economic profitability. The return and risk measures of long-only trading strategies indicate that machine learning models produce size-able gains relative to our benchmark. Neural networks perform best, even after adjusting for risk and accounting for transaction costs. However, a classification-based portfolio formation, utilizing a support vector machine that avoids estimating stock-level expected returns, performs even better than the neural network architecture.

Du, Y., Wang, J., Feng, W., Pan, S., Qin, T., Xu, R., and Wang, C. (2021). "AdaRNN: Adaptive Learning and Forecasting of Time Series." In: *arXiv e-Print*.

Time series has wide applications in the real world and is known to be difficult to forecast. Since its statistical properties change over time, its distribution also changes temporally, which will cause severe distribution shift problem to existing methods. However, it remains unexplored to model the time series in the distribution perspective. In this paper, we term this as Temporal Covariate Shift (TCS). This paper proposes Adaptive RNNs (AdaRNN) to tackle the TCS problem by building an adaptive model that generalizes well on the unseen test data. AdaRNN is sequentially composed of two novel algorithms. First, we propose Temporal Distribution Characterization to better characterize the distribution information in the TS. Second, we propose Temporal Distribution Matching to reduce the distribution mismatch in TS to learn the adaptive TS model. AdaRNN is a general framework with flexible distribution distances integrated. Experiments on human activity recognition, air quality prediction, and financial analysis show that AdaRNN outperforms the latest methods by a classification accuracy of 2.6% and significantly reduces the RMSE by 9.0%. We also show that the temporal distribution matching algorithm can be extended in Transformer structure to boost its performance.

Elkamhi, R., Lee, J. S. H., and Salerno, M. (2021). "Factor Investing Using Capital Market Assumptions." In: *The Journal of Portfolio Management*.

Capital market assumptions (CMAs), which are long-term risk and return forecasts for asset classes, are important pillars of the investment industry. However, applying them reliably in portfolio construction has been (and still is) a challenge in the industry. This article demonstrates that, despite the difficulties, CMAs are useful for building an investment portfolio using a factor approach. Using a small set of macroeconomic factors, the authors detail a methodology for deriving a factor model from CMAs and then use it to show that (1) these factors price the expected returns from CMAs and (2) the mean-variance factor allocations are substantially more stable than the mean-variance asset portfolios. Furthermore, this article outlines a new approach to building an asset portfolio that respects a desired factor allocation. Overall, this article helps reduce the barrier to entry for factor-based portfolio construction by providing a recipe for building factor models and performing factor-based portfolio construction using publicly available CMAs.

Ellingsen, J., Larsen, V. H., and Thorsrud, L. A. (2022). "News media vs. FRED-MD for macroeconomic forecasting." In: *Journal of Applied Econometrics*.

Using a unique dataset of 22.5 million news articles from the Dow Jones Newswires Archive, we perform an in depth real-time out-of-sample forecasting comparison study with one of the most widely used data sets in the newer forecasting literature, namely the FRED-MD dataset. Focusing on U.S. GDP, consumption and investment growth, our results suggest that the news data contains information not captured by the hard economic indicators, and that the news-based data are particularly informative for forecasting consumption developments.

Faloutsos, C., Flunkert, V., Gasthaus, J., Januschowski, T., and Wang, Y. (2019). "Forecasting Big Time Series: Theory and Practice." In: *Tutorial for the 25TH ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM.

Time series forecasting is a key ingredient in the automation and optimization of business processes: in retail, deciding which products to order and where to store them depends on the forecasts of future demand in different regions; in cloud computing, the estimated future usage of services and infrastructure components guides capacity planning; and workforce scheduling in warehouses and factories requires forecasts of the future workload. Recent years have witnessed a paradigm shift in forecasting techniques and applications, from computer-assisted model- and assumption-based to data-driven and fully-automated. This shift can be attributed to the availability of large, rich, and diverse time series data sources and result in a set of challenges that need to be addressed, such as the following: How can we build statistical models to efficiently and effectively learn to forecast from large and diverse data sources? How can we leverage the statistical power of "similar" time series to improve forecasts in the case of limited observations? What are the implications for building forecasting systems that can handle large data volumes? The objective of this tutorial is to provide a concise and intuitive overview of the most important methods and tools available for solving large-scale forecasting problems. We review the state of the art in three related fields: (1) classical modeling of time series, (2) modern methods including tensor analysis and deep learning for forecasting. Furthermore, we discuss the practical aspects of building a large scale forecasting system, including data integration, feature generation, backtest framework, error tracking and analysis, etc. While our focus is on providing an intuitive overview of the methods and practical issues which we will illustrate via case studies and interactive materials with Jupyter notebooks.

Fama, E. F. and French, K. R. (2018). "Choosing factors." In: *Journal of Financial Economics* 128(2), pp. 234–252.

Our goal is to develop insights about the maximum squared Sharpe ratio for model factors as a metric for ranking asset pricing models. We consider nested and non-nested models. The nested models are the capital asset pricing model, the three-factor model of Fama and French (1993), the five-factor extension in Fama and French (2015), and a six-factor model that adds a momentum factor. The non-nested models examine three issues about factor choice in the six-factor model: (1) cash profitability versus operating profitability as the variable used to construct profitability factors, (2) long-short spread factors versus excess return factors, and (3) factors that use small or big stocks versus factors that use both.

Fama, E. F. and French, K. R. (2020). "Comparing Cross-Section and Time-Series Factor Models." In: *The Review of Financial Studies* 33(5), pp. 1891–1926.

We use the cross-section regression approach of Fama and MacBeth (1973) to construct cross-section factors corresponding to the time-series factors of Fama and French (2015). Time-series models that use only cross-section factors provide better descriptions of average returns than time-series models that use time-series factors. This is true when we impose constant factor loadings and when we use time-varying loadings that are natural for time-series factors and time-varying loadings that are natural for cross-section factors.

Fameliti, S. P. and Skintzi, V. D. (2020). "Predictive ability and economic gains from volatility forecast combinations." In: *Journal of Forecasting* 39(2), pp. 200–219.

The availability of numerous modeling approaches for volatility forecasting leads to model uncertainty for both researchers and practitioners. A large number of studies provide evidence in favor of combination methods for forecasting a variety of financial variables, but most of them are implemented on returns forecasting and evaluate their performance based solely on statistical evaluation criteria. In this paper, we combine various volatility forecasts based on different combination schemes and evaluate their performance in forecasting the volatility of the S&P 500 index. We use an exhaustive variety of combination methods to forecast volatility, ranging from simple techniques to time-varying techniques based on the past performance of the single models and regression techniques. We then evaluate the forecasting performance of single and combination volatility forecasts based on both statistical and economic loss functions. The empirical analysis in this paper yields an important conclusion. Although combination forecasts based on more complex methods perform better than the simple combinations and single models, there is no dominant combination technique that outperforms the rest in both statistical and economic terms.

Fan, J., Ke, Y., Sun, Q., and Zhou, W.-X. (2019). "FarmTest: Factor-adjusted robust multiple testing with approximate false discovery control." In: *Journal of the American Statistical Association* 114(528), pp. 1880–1893.

Large-scale multiple testing with correlated and heavy-tailed data arises in a wide range of research areas from genomics, medical imaging to finance. Conventional methods for estimating the false discovery proportion (FDP) often ignore the effect of heavy-tailedness and the dependence structure among test statistics, and thus may lead to inefficient or even inconsistent estimation. Also, the commonly imposed joint normality assumption is arguably too stringent for many applications. To address these challenges, in this paper we propose a Factor-Adjusted Robust Multiple Testing (FarmTest) procedure for large-scale simultaneous inference with control of the false discovery proportion. We demonstrate that robust factor adjustments are extremely important in both controlling the FDP and improving the power. We identify general conditions under which the proposed method produces consistent estimate of the FDP. As a byproduct that is of independent interest, we establish an exponential-type deviation inequality for a robust U-type covariance estimator under the spectral norm. Extensive numerical experiments demonstrate the advantage of the proposed method over several state-of-the-art methods especially when the data are generated from heavy-tailed distributions. The proposed procedures are implemented in the R-package FarmTest.

Fang, Y., Guan, B., Wu, S., and Heravi, S. (2020). "Optimal forecast combination based on ensemble empirical mode decomposition for agricultural commodity futures prices." In: *Journal of Forecasting* 39(6), pp. 877–886.

Improving the prediction accuracy of agricultural product futures prices is important for investors, agricultural producers, and policymakers. This is to evade risks and enable government departments to formulate appropriate agricultural regulations and policies. This study employs the ensemble empirical mode decomposition (EEMD) technique to decompose six different categories of agricultural futures prices. Subsequently, three models – support vector machine (SVM), neural network (NN), and autoregressive integrated moving average (ARIMA) –are used to predict the decomposition components. The final hybrid model is then constructed by comparing the prediction performance of the decomposition components. The predicting performance of the combination model is then compared with the benchmark individual models: SVM, NN, and ARIMA. Our main interest in this study is on short-term forecasting, and thus we only consider 1-day and 3-day forecast horizons. The results indicate

that the prediction performance of the EEMD combined model is better than that of individual models, especially for the 3-day forecasting horizon. The study also concluded that the machine learning methods outperform the statistical methods in forecasting high-frequency volatile components. However, there is no obvious difference between individual models in predicting low-frequency components.

Faria, G. and Verona, F. (2021). "Time-frequency forecast of the equity premium." In: *Quantitative Finance*.
Any time series can be decomposed into cyclical components fluctuating at different frequencies. Accordingly, in this paper, we propose a method to forecast the equity premium which exploits the frequency relationship between the equity premium and several predictor variables. We evaluate a large set of models and find that, by selecting the relevant frequencies for equity premium forecasting purposes, this method significantly improves in a statistical and economic way upon standard time series forecasting methods. This outperformance is robust regardless of the predictor used, the out-of-sample period considered, and the frequency of the data used.

Fauvel, K., Masson, V., and Fromont, elisa (2021). "A Performance-Explainability Framework to Benchmark Machine Learning Methods: Application to Multivariate Time Series Classifiers." In: *arXiv e-Print*.
Our research aims to propose a new performance-explainability analytical framework to assess and benchmark machine learning methods. The framework details a set of characteristics that systematize the performance-explainability assessment of existing machine learning methods. In order to illustrate the use of the framework, we apply it to benchmark the current state-of-the-art multivariate time series classifiers.

Feldkircher, M., Huber, F., and Pfarrhofer, M. (2020). "Factor Augmented Vector Autoregressions, Panel VARs, and Global VARs." In: *Macroeconomic Forecasting in the Era of Big Data*. Springer International Publishing, pp. 65–93.
This chapter provides a thorough introduction to panel, global, and factor augmented vector autoregressive models. These models are typically used to capture interactions across units (i.e., countries) and variable types. Since including a large number of countries and/or variables increases the dimension of the models, all three approaches aim to decrease the dimensionality of the parameter space. After introducing each model, we briefly discuss key specification issues. A running toy example serves to highlight this point and outlines key differences across the different models. To illustrate the merits of the competing approaches, we perform a forecasting exercise and show that it pays off to introduce cross-sectional information in terms of forecasting key macroeconomic quantities.

Feng, G., Giglio, S., and Xiu, D. (2020). "Taming the factor zoo: A test of new factors." In: *The Journal of Finance* 75(3), pp. 1327–1370.
We propose a model selection method to systematically evaluate the contribution to asset pricing of any new factor, above and beyond what a high-dimensional set of existing factors explains. Our methodology accounts for model selection mistakes that produce a bias due to omitted variables, unlike standard approaches that assume perfect variable selection. We apply our procedure to a set of factors recently discovered in the literature. While most of these new factors are shown to be redundant relative to the existing factors, a few have statistically significant explanatory power beyond the hundreds of factors proposed in the past.

Filipovic, D. and Khalilzadeh, A. (2021). "Machine Learning for Predicting Stock Return Volatility." In: *SSRN e-Print*.
We use machine learning methods to predict stock return volatility. Our out-of-sample prediction of realised volatility for a large cross-section of US stocks over the sample period from 1992 to 2016 is on average 44.1% against the actual realised volatility of 43.8% with an R2 being as high as double the ones reported in the literature. We further show that machine learning methods can capture the stylized facts about volatility without relying on any assumption about the distribution of stock returns. Finally, we show that our long short-term memory model outperforms other models by properly carrying information from the past predictor values.

Fjellstrom, C. (2022). "Long Short-Term Memory Neural Network for Financial Time Series." In: *arXiv e-Print*.
Performance forecasting is an age-old problem in economics and finance. Recently, developments in machine learning and neural networks have given rise to non-linear time series models that provide modern and promising alternatives to traditional methods of analysis. In this paper, we present an ensemble of independent and parallel long short-term memory (LSTM) neural networks for the prediction of stock price movement. LSTMs have been shown to be especially suited for time series data due to their ability to incorporate past information, while neural network ensembles have been found to reduce variability in results and improve generalization. A binary classification problem based on the median of returns is used, and the ensemble's forecast depends on a threshold value, which is the minimum number of LSTMs required to agree upon the result. The model is applied to the constituents of the smaller, less efficient Stockholm OMX30 instead of other major market indices such as the

DJIA and S&P500 commonly found in literature. With a straightforward trading strategy, comparisons with a randomly chosen portfolio and a portfolio containing all the stocks in the index show that the portfolio resulting from the LSTM ensemble provides better average daily returns and higher cumulative returns over time. Moreover, the LSTM portfolio also exhibits less volatility, leading to higher risk-return ratios.

Fleiss, A. and Cui, H. (2021). "Forecasting Stock Price Changes Using Natural Language Processing." In: *SSRN e-Print*.

All investors attempt to predict stock market returns when they make investment decisions. However, making such predictions is not a trivial task. As a result, many strategies have been proposed by researchers as potential ways to predict stock returns. More recently, data analytics - in general - and natural language processing, in particular, have been identified as viable options. In this project, we investigate the use of natural language processing to forecast stock price changes. Specifically, we analyze firms' 10-K and 10-Q reports to identify sentiment. Using the computed sentiment scores, we develop models to predict the direction of stock price movements both in the short run and in the long run. Our first step in developing these models is to investigate some sentiment scoring methods and apply them to the Loughran-McDonald dictionary. Next, we use the model word2vec to extend the usage of the Loughran-McDonald dictionary and then apply the sentiment metrics. Additionally, we apply the proposed sentiment metrics to FinBERT, which learns contextual relations between words. Finally, we build supervised machine learning algorithms that use the proposed sentiments as inputs to forecast price changes. We train our algorithms on 10-K and 10-Q reports of 48 companies in the S&P 500 from 2013 to 2017. Finally, we test our models on the corresponding reports from 2018 to 2019 and conclude that predictive signals can be extracted from 10-K and 10-Q reports.

Fosten, J. and Gutknecht, D. (2021). "Horizon confidence sets." In: *Empirical Economics*.

This paper introduces a new statistical procedure to discriminate between competing forecasting models at different forecast horizons. Unlike existing tests, which eliminate a model from all horizons if dominated according to some loss measure, our methodology identifies an "optimal" set of models at each horizon, retaining a model that performs well at a given horizon even if dominated at others. While our method is especially useful in applications to long-term forecasting as well as short-term nowcasting, it can also be applied in wider settings like the comparison of forecasting models across countries. We conduct a small Monte Carlo study to investigate the finite sample properties and apply our procedure to nowcasting US real GDP growth and its subcomponents, comparing a factor-based nowcasting method to a naive benchmark. Unlike existing methods, ours can formally detect the point in the quarter at which the factor method beats the benchmark or vice versa.

Frenkel, M., Jung, J.-K., and Rulke, J.-C. (2021). "Testing for the rationality of central bank interest rate forecasts." In: *Empirical Economics*.

In this paper, we study the bias in interest rate projections of five central banks, namely the central banks of the Czech Republic, New Zealand, Norway, Sweden, and the USA. We examine whether central bank projections are based on an asymmetric loss function and report evidence that central banks perceive an overprojection of their longer-term interest rate forecasts as twice as costly as an underprojection of the same size. We find that forecast rationality is consistent with biased interest rate projections under the assumption of an asymmetric loss function, which contributes to explaining the behavior of the examined central banks and their forecasts.

Freyberger, J., Neuhierl, A., and Weber, M. (2020). "Dissecting Characteristics Nonparametrically." In: *The Review of Financial Studies* 33(5), pp. 2326–2377.

We propose a nonparametric method to study which characteristics provide incremental information for the cross-section of expected returns. We use the adaptive group LASSO to select characteristics and to estimate how selected characteristics affect expected returns nonparametrically. Our method can handle a large number of characteristics and allows for a flexible functional form. Our implementation is insensitive to outliers. Many of the previously identified return predictors don't provide incremental information for expected returns, and nonlinearities are important. We study our method's properties in simulations and find large improvements in both model selection and prediction compared to alternative selection methods. Authors have furnished an Internet Appendix, which is available on the Oxford University Press Web site next to the link to the final published paper online.

Fulton, C. and Hubrich, K. (2021). "Forecasting US Inflation in Real Time." In: *SSRN e-Print* (014).

We perform a real-time forecasting exercise for US inflation, investigating whether and how additional information–additional macroeconomic variables, expert judgment, or forecast combination–can improve forecast accuracy and robustness. In our analysis we consider the pre-pandemic period including the Global Financial Crisis and the following expansion–the longest on record–featuring unemployment that fell to a rate not seen for nearly

sixty years. Distinguishing features of our study include the use of published Federal Reserve Board staff forecasts contained in Tealbooks and a focus on forecasting performance before, during, and after the Global Financial Crisis, with relevance also for the current crisis and beyond.

We find that while simple models remain hard to beat, the additional information that we consider can improve forecasts, especially in the post-crisis period. Our results show that

1) forecast combination approaches improve forecast accuracy over simpler models and robustify against bad forecasts, a particularly relevant feature in the current environment;
2) aggregating forecasts of inflation components can improve performance compared to forecasting the aggregate directly;
3) judgmental forecasts, which likely incorporate larger and more timely datasets, provide improved forecasts at short horizons.

.

Gafka, B., Savor, P. G., and Wilson, M. I. (2021). "Sources of Return Predictability." In: *SSRN e-Print*.
A large literature establishes a set of predictors that robustly forecast future market returns, raising questions about these predictors' origins. We develop an approach to determine whether a particular predictor represents a proxy for fundamental risk, which is based on an intuitive assumption that risk-based predictors should be linked to new information about economic conditions. We show that each predictor forecasts returns either on days with macroeconomic announcements or on other days, but never on both types of days. These results suggest that sources of return predictability differ across predictors, with some driven by economic fundamentals and others having different origins. Consistent with this interpretation, announcement-day returns are positively related to future changes in fundamental value, while there is no such relation for non-announcement-day returns.

Gasthaus, J., Benidis, K., Wang, Y., Rangapuram, S. S., Salinas, D., Flunkert, V., and Januschowski, T. (2019). "Probabilistic Forecasting with Spline Quantile Function RNNs." In: *Proceedings of Machine Learning Research* 89, pp. 1901–1910.
In this paper, we propose a flexible method for probabilistic modeling with conditional quantile functions using monotonic regression splines. The shape of the spline is parameterized by a neural network whose parameters are learned by minimizing the continuous ranked probability score. Within this framework, we propose a method for probabilistic time series forecasting, which combines the modeling capacity of recurrent neural networks with the flexibility of a spline-based representation of the output distribution. Unlike methods based on parametric probability density functions and maximum likelihood estimation, the proposed method can flexibly adapt to different output distributions without manual intervention. We empirically demonstrate the effectiveness of the approach on synthetic and real-world data sets.

Gastinger, J., Nicolas, S., Stepic, D., Schmidt, M., and Schulke, A. (2021). "A study on Ensemble Learning for Time Series Forecasting and the need for Meta-Learning." In: *arXiv e-Print*.
The contribution of this work is twofold: (1) We introduce a collection of ensemble methods for time series forecasting to combine predictions from base models. We demonstrate insights on the power of ensemble learning for forecasting, showing experiment results on about 16000 openly available datasets, from M4, M5, M3 competitions, as well as FRED (Federal Reserve Economic Data) datasets. Whereas experiments show that ensembles provide a benefit on forecasting results, there is no clear winning ensemble strategy (plus hyperparameter configuration). Thus, in addition, (2), we propose a meta-learning step to choose, for each dataset, the most appropriate ensemble method and their hyperparameter configuration to run based on dataset meta-features.

Geertsema, P. G. and Lu, H. (2021). "Long-horizon predictability and information decay in equity markets." In: *SSRN e-Print*.
We use machine learning to predict stock returns at forward horizons from 1 month ahead to 120 months ahead. Stock return predictability declines with the forecast horizon; it follows an asymptotic exponential decay process consisting of a permanent component (c. 20 bp/month) and a transient component (c. 240 bp/month) which decays at around 6% per month. Persistent but declining predictability at increasing horizons can be explained by persistent but declining benchmark risk factor exposures at each horizon. Limits to arbitrage linked to risk (but not implementation costs) also explains declining profitability.

Geweke, J. and Amisano, G. (2010). "Comparing and evaluating Bayesian predictive distributions of asset returns." In: *International Journal of Forecasting* 26(2), pp. 216–230.

Bayesian inference in a time series model provides exact out-of-sample predictive distributions that fully and coherently incorporate parameter uncertainty. This study compares and evaluates Bayesian predictive distributions from alternative models, using as an illustration five alternative models of asset returns applied to daily SandP 500 returns from the period 1976 through 2005. The comparison exercise uses predictive likelihoods and is inherently Bayesian. The evaluation exercise uses the probability integral transformation and is inherently frequentist. The illustration shows that the two approaches can be complementary, with each identifying strengths and weaknesses in models that are not evident using the other.

Ghorbani, M. and Chong, E. K. P. (2021). "A dimension reduction method for stock-price prediction using multiple predictors." In: *Operational Research*.

Stock-price prediction has been the focus of extensive studies. Historical price values have been proven to have power to predict future prices. At the same time, different economic variables have also been used in the literature to predict stock-price values with high accuracy. In this work, we develop a general method for stock-price prediction using multiple predictors. First, we use multichannel cross-correlation coefficient as a measure for selecting the most correlated set of variables for each stock. We then construct the temporally local covariance matrix of the data and use this as the basis for a dimension-reduction method for prediction. This method involves resolving the predictive data (predictors) onto a principal subspace and from there producing a prediction that is consistent with the resolved data. Our method is easily implemented and can accommodate an arbitrary number of predictors. We investigate the optimal number of predictors based on two performance metrics: mean squared error of the prediction and the directional change statistic. We illustrate our results based on historical daily price data for 50 companies.

Gilleland, E., Hering, A. S., Fowler, T. L., and Brown, B. G. (2018). "Testing the Tests: What Are the Impacts of Incorrect Assumptions When Applying Confidence Intervals or Hypothesis Tests to Compare Competing Forecasts?" In: *Monthly Weather Review* 146(6), pp. 1685–1703.

Which of two competing continuous forecasts is better? This question is often asked in forecast verification, as well as climate model evaluation. Traditional statistical tests seem to be well suited to the task of providing an answer. However, most such tests do not account for some of the special underlying circumstances that are prevalent in this domain. For example, model output is seldom independent in time, and the models being compared are geared to predicting the same state of the atmosphere, and thus they could be contemporaneously correlated with each other. These types of violations of the assumptions of independence required for most statistical tests can greatly impact the accuracy and power of these tests. Here, this effect is examined on simulated series for many common testing procedures, including two-sample and paired t and normal approximation z tests, the z test with a first-order variance inflation factor applied, and the newer Hering-Genton (HG) test, as well as several bootstrap methods. While it is known how most of these tests will behave in the face of temporal dependence, it is less clear how contemporaneous correlation will affect them. Moreover, it is worthwhile knowing just how badly the tests can fail so that if they are applied, reasonable conclusions can be drawn. It is found that the HG test is the most robust to both temporal dependence and contemporaneous correlation, as well as the specific type and strength of temporal dependence. Bootstrap procedures that account for temporal dependence stand up well to contemporaneous correlation and temporal dependence, but require large sample sizes to be accurate.

Gilliland, M. (2020). "The value added by machine learning approaches in forecasting." In: *International Journal of Forecasting* 36(1), pp. 161–166.

This discussion reflects on the results of the M4 forecasting competition, and in particular, the impact of machine learning (ML) methods. Unlike the M3, which included only one ML method (an automatic artificial neural network that performed poorly), M4 49 participants included eight that used either pure ML approaches, or ML in conjunction with statistical methods. The six pure (or combination of pure) ML methods again fared poorly, with all of them falling below the Comb benchmark that combined three simple time series methods. However, utilizing ML either in combination with statistical methods (and for selecting weightings) or in a hybrid model with exponential smoothing not only exceeded the benchmark, but performed at the top. While these promising results by no means prove ML to be a panacea, they do challenge the notion that complex methods do not add value to the forecasting process.

Giovannelli, A., Massacci, D., and Soccorsi, S. (2021a). "Forecasting Stock Returns with Large Dimensional Factor Models." In: *SSRN e-Print*.

Assuming that economic variables can be decomposed into common and idiosyncratic components, we study equity premium out-of-sample predictability when the information contained in a high number of predictors is extracted using large dimensional factor models. We consider factor models with a static representation of the common components, and models that allow for dynamic and infinite-dimensional representations. Using statistical and economic evaluation criteria, we show that large dimensional factor models with a dynamic representation of the common components do help predicting the equity premium. Furthermore, exploiting the well-known link between the business cycle and return predictability, we find more accurate predictions by combining rolling and recursive forecasts.

Giovannelli, A., Massacci, D., and Soccorsi, S. (2021b). "Forecasting stock returns with large dimensional factor models." In: *Journal of Empirical Finance* 63, pp. 252–269.

We study equity premium out-of-sample predictability by extracting the information contained in a high number of macroeconomic predictors via large dimensional factor models. We compare the well-known factor model with a static representation of the common components with the Generalized Dynamic Factor Model, which accounts for time series dependence in the common components. Using statistical and economic evaluation criteria, we empirically show that the Generalized Dynamic Factor Model helps predicting the equity premium. Exploiting the link between business cycle and return predictability, we find accurate predictions also by combining rolling and recursive forecasts in real-time.

Gneiting, T. and Resin, J. (2022). "Regression Diagnostics meets Forecast Evaluation: Conditional Calibration, Reliability Diagrams, and Coefficient of Determination." In: *arXiv e-Print*.

Model diagnostics and forecast evaluation are two sides of the same coin. A common principle is that fitted or predicted distributions ought to be calibrated or reliable, ideally in the sense of auto-calibration, where the outcome is a random draw from the posited distribution. For binary responses, this is the universal concept of reliability. For real-valued outcomes, a general theory of calibration has been elusive, despite a recent surge of interest in distributional regression and machine learning. We develop a framework rooted in probability theory, which gives rise to hierarchies of calibration, and applies to both predictive distributions and stand-alone point forecasts. In a nutshell, a prediction - distributional or single-valued - is conditionally T-calibrated if it can be taken at face value in terms of the functional T. Whenever T is defined via an identification function - as in the cases of threshold (non) exceedance probabilities, quantiles, expectiles, and moments - auto-calibration implies T-calibration. We introduce population versions of T-reliability diagrams and revisit a score decomposition into measures of miscalibration (MCB), discrimination (DSC), and uncertainty (UNC). In empirical settings, stable and efficient estimators of T-reliability diagrams and score components arise via nonparametric isotonic regression and the pool-adjacent-violators algorithm. For in-sample model diagnostics, we propose a universal coefficient of determination, $R^* = \dfrac{DSC - MCB}{UNC}$ that nests and reinterprets the classical R^2 in least squares (mean) regression and its natural analogue R^1 in quantile regression, yet applies to T-regression in general, with MCB ≥ 0, DSC ≥ 0, and $R^* \in [0, 1]$ under modest conditions.

Godahewa, R., Bandara, K., Webb, G. I., Smyl, S., and Bergmeir, C. (2021a). "Ensembles of localised models for time series forecasting." In: *Knowledge-Based Systems* 233, p. 107518.

With large quantities of data typically available nowadays, forecasting models that are trained across sets of time series, known as Global Forecasting Models (GFM), are regularly outperforming traditional univariate forecasting models that work on isolated series. As GFMs usually share the same set of parameters across all time series, they often have the problem of not being localised enough to a particular series, especially in situations where datasets are heterogeneous. We study how ensembling techniques can be used with generic GFMs and univariate models to solve this issue. Our work systematises and compares relevant current approaches, namely clustering series and training separate submodels per cluster, the so-called ensemble of specialists approach, and building heterogeneous ensembles of global and local models. We fill some gaps in the existing GFM localisation approaches, in particular by incorporating varied clustering techniques such as feature-based clustering, distance-based clustering and random clustering, and generalise them to use different underlying GFM model types. We then propose a new methodology of clustered ensembles where we train multiple GFMs on different clusters of series, obtained by changing the number of clusters and cluster seeds. Using Feed-forward Neural Networks, Recurrent Neural Networks, and Pooled Regression models as the underlying GFMs, in our evaluation on eight publicly available datasets, the proposed models are able to achieve significantly higher accuracy than baseline GFM models and univariate forecasting methods.

Godahewa, R., Bergmeir, C., Webb, G. I., Hyndman, R. J., and Montero-Manso, P. (2021b). "Monash Time Series Forecasting Archive." In: *arXiv E-Print*.

Many businesses and industries nowadays rely on large quantities of time series data making time series forecasting an important research area. Global forecasting models that are trained across sets of time series have shown

a huge potential in providing accurate forecasts compared with the traditional univariate forecasting models that work on isolated series. However, there are currently no comprehensive time series archives for forecasting that contain datasets of time series from similar sources available for the research community to evaluate the performance of new global forecasting algorithms over a wide variety of datasets. In this paper, we present such a comprehensive time series forecasting archive containing 20 publicly available time series datasets from varied domains, with different characteristics in terms of frequency, series lengths, and inclusion of missing values. We also characterise the datasets, and identify similarities and differences among them, by conducting a feature analysis. Furthermore, we present the performance of a set of standard baseline forecasting methods over all datasets across eight error metrics, for the benefit of researchers using the archive to benchmark their forecasting algorithms.

Goliński, A. and Spencer, P. (2021). "Estimating the Term Structure with Linear Regressions: Getting to the Roots of the Problem." In: *Journal of Financial Econometrics* 19(5), pp. 960–984.

Linear estimators of the affine term structure model are inconsistent since they cannot reproduce the factors used in estimation. This is a serious handicap empirically, giving a worse fit than the conventional ML estimator that ensures consistency. We show that a simple self-consistent estimator can be constructed using the eigenvalue decomposition of a regression estimator. The remaining parameters of the model follow analytically. Estimates from this model are virtually indistinguishable from that of the ML estimator. We apply the method to estimate various models of U.S. Treasury yields. These exercises greatly extend the range of models that can be estimated.

Golyandina, N., Korobeynikov, A., and Zhigljavsky, A. (2018). *Singular Spectrum Analysis with R*. Springer Berlin Heidelberg. 272 pp.

This comprehensive and richly illustrated volume provides up-to-date material on Singular Spectrum Analysis (SSA). SSA is a well-known methodology for the analysis and forecasting of time series. Since quite recently, SSA is also being used to analyze digital images and other objects that are not necessarily of planar or rectangular form and may contain gaps. SSA is multi-purpose and naturally combines both model-free and parametric techniques, which makes it a very special and attractive methodology for solving a wide range of problems arising in diverse areas, most notably those associated with time series and digital images. An effective, comfortable and accessible implementation of SSA is provided by the R-package Rssa, which is available from CRAN and reviewed in this book. Written by prominent statisticians who have extensive experience with SSA, the book (a) presents the up-to-date SSA methodology, including multidimensional extensions, in language accessible to a large circle of users, (b) combines different versions of SSA into a single tool, (c) shows the diverse tasks that SSA can be used for, (d) formally describes the main SSA methods and algorithms, and (e) provides tutorials on the Rssa package and the use of SSA. The book offers a valuable resource for a very wide readership, including professional statisticians, specialists in signal and image processing, as well as specialists in numerous applied disciplines interested in using statistical methods for time series analysis, forecasting, signal and image processing. The book is written on a level accessible to a broad audience and includes a wealth of examples; hence it can also be used as a textbook for undergraduate and postgraduate courses on time series analysis and signal processing.

Gonzalez-Rivera, G., Luo, Y., and Ruiz, E. (2020). "Prediction regions for interval-valued time series." In: *Journal of Applied Econometrics* 35(4), pp. 373–390.

We approximate probabilistic forecasts for interval-valued time series by offering alternative approaches. After fitting a possibly non-Gaussian bivariate vector autoregression (VAR) model to the center/log-range system, we transform prediction regions (analytical and bootstrap) for this system into regions for center/range and upper/lower bounds systems. Monte Carlo simulations show that bootstrap methods are preferred according to several new metrics. For daily S&P 500 low/high returns, we build joint conditional prediction regions of the return level and volatility. We illustrate the usefulness of obtaining bootstrap forecasts regions for low/high returns by developing a trading strategy and showing its profitability when compared to using point forecasts.

Gospodinov, N. and Maasoumi, E. (2021). "Generalized aggregation of misspecified models: with an application to asset pricing." In: *Journal of Econometrics* 222(1), pp. 451–467.

We propose a generalized aggregation approach for model averaging. The entropy-based optimal criterion is a natural choice for aggregating information from many ”globally” misspecified models as it adapts better to the underlying model uncertainty and obtains more robust approximations. Unlike almost all other approaches in the existing literature, we do not require a ”reference model,” or a true data generation process contained in the set of models - neither implicitly nor in otherwise popular limiting forms. This shift in paradigm prioritizes stochastic optimization and aggregation of information about outcomes over parameter estimation of an optimally selected model. Stochastic optimization is based on a risk function of aggregators across models that satisfies oracle

inequalities. Our generalized aggregators relax the common perfect substitutability of the candidate models, implicit in linear averaging and pooling. The aggregation weights are data-driven and obtained from a proper (Hellinger) distance measure. The empirical results illustrate the performance and economic significance of the aggregation approach in the context of stochastic discount factor models and inflation forecasting.

Gospodinov, N. and Robotti, C. (2021). "Common pricing across asset classes: Empirical evidence revisited." In: *Journal of Financial Economics* 132(2), pp. 292–324.
Intermediary and downside risk asset pricing theories lay the foundations for spanning the multi-asset return space by a small number of risk factors. Recent studies show strong empirical support for such factors across major asset classes. We revisit these results and show that robust evidence for common factor pricing remains elusive. Importantly, the proposed risk factors do not seem to provide incremental information to the traditional market factor. We argue that most of the economic and statistical challenges are not specific to these analyses and, with the aid of a placebo test, offer general recommendations for improving empirical practice, thus adding to the prescriptions in Lewellen et al. (2010).

Goyal, A. and Jegadeesh, N. (2018). "Cross-Sectional and Time-Series Tests of Return Predictability: What Is the Difference?" In: *The Review of Financial Studies* 31(5), pp. 1784–1824.
We compare the performance of time-series (TS) and cross-sectional (CS) strategies based on past returns. While CS strategies are zero-net investment long/short strategies, TS strategies take on a time-varying net long investment in risky assets. For individual stocks, the difference between the performances of TS and CS strategies is largely due to this time-varying net long investment. With multiple international asset classes with heterogeneous return distributions, scaled CS strategies significantly outperform similarly scaled TS strategies.

Gramespacher, T. and Banziger, A. (2019). "The Bias in Two-Pass Regression Tests of Asset-Pricing Models in Presence of Idiosyncratic Errors with Cross-Sectional Dependence." In: *Review of Pacific Basin Financial Markets and Policies* 22(02), p. 1950012.
In two-pass regression-tests of asset-pricing models, cross-sectional correlations in the errors of the first-pass time-series regression lead to correlated measurement errors in the betas used as explanatory variables in the second-pass cross-sectional regression. The slope estimator of the second-pass regression is an estimate for the factor risk-premium and its significance is decisive for the validity of the pricing model. While it is well known that the slope estimator is downward biased in presence of uncorrelated measurement errors, we show in this paper that the correlations seen in empirical return data substantially suppress this bias. For the case of a single-factor model, we calculate the bias of the OLS slope estimator in the presence of correlated measurement errors with a first-order Taylor-approximation in the size of the errors. We show that the bias increases with the size of the errors, but decreases the more the errors are correlated. We illustrate and validate our result using a simulation approach based on empirical data commonly used in asset-pricing tests.

Graziani, C., Rosner, R., Adams, J. M., and Machete, R. L. (2021). "Probabilistic recalibration of forecasts." In: *International Journal of Forecasting* 37(1), pp. 1–27.
We present a scheme by which a probabilistic forecasting system whose predictions have a poor probabilistic calibration may be recalibrated through the incorporation of past performance information in order to produce a new forecasting system that is demonstrably superior to the original, inasmuch as one may use it to win wagers consistently against someone who is using the original system. The scheme utilizes Gaussian process (GP) modeling to estimate a probability distribution over the probability integral transform (PIT) of a scalar predictand. The GP density estimate gives closed-form access to information entropy measures that are associated with the estimated distribution, which allows the prediction of winnings in wagers against the base forecasting system. A separate consequence of the procedure is that the recalibrated forecast has a uniform expected PIT distribution. One distinguishing feature of the procedure is that it is appropriate even if the PIT values are not i.i.d. The recalibration scheme is formulated in a framework that exploits the deep connections among information theory, forecasting, and betting. We demonstrate the effectiveness of the scheme in two case studies: a laboratory experiment with a nonlinear circuit and seasonal forecasts of the intensity of the El Nino-Southern Oscillation phenomenon.

Grazzi, R., Flunkert, V., Salinas, D., Januschowski, T., Seeger, M., and Archambeau, C. (2021). "Meta-Forecasting by combining Global Deep Representations with Local Adaptation." In: *arXiv e-Print*.
While classical time series forecasting considers individual time series in isolation, recent advances based on deep learning showed that jointly learning from a large pool of related time series can boost the forecasting accuracy. However, the accuracy of these methods suffers greatly when modeling out-of-sample time series, significantly limiting their applicability compared to classical forecasting methods. To bridge this gap, we adopt a meta-

learning view of the time series forecasting problem. We introduce a novel forecasting method, called Meta Global-Local Auto-Regression (Meta-GLAR), that adapts to each time series by learning in closed-form the mapping from the representations produced by a recurrent neural network (RNN) to one-step-ahead forecasts. Crucially, the parameters ofthe RNN are learned across multiple time series by backpropagating through the closed-form adaptation mechanism. In our extensive empirical evaluation we show that our method is competitive with the state-of-the-art in out-of-sample forecasting accuracy reported in earlier work.

Greenberg, S. (2020). "Calibration Scoring Rules for Practical Prediction Training." In: *arXiv e-Print*.

In situations where forecasters are scored on the quality of their probabilistic predictions, it is standard to use 'proper' scoring rules to perform such scoring. These rules are desirable because they give forecasters no incentive to lie about their probabilistic beliefs. However, in the real world context of creating a training program designed to help people improve calibration through prediction practice, there are a variety of desirable traits for scoring rules that go beyond properness. These potentially may have a substantial impact on the user experience, usability of the program, or efficiency of learning. The space of proper scoring rules is too broad, in the sense that most proper scoring rules lack these other desirable properties. On the other hand, the space of proper scoring rules is potentially also too narrow, in the sense that we may sometimes choose to give up properness when it conflicts with other properties that are even more desirable from the point of view of usability and effective training. We introduce a class of scoring rules that we call 'Practical' scoring rules, designed to be intuitive to users in the context of 'right' vs. 'wrong' probabilistic predictions. We also introduce two specific scoring rules for prediction intervals, the 'Distance' and 'Order of magnitude' rules. These rules are designed to satisfy a variety of properties that, based on user testing, we believe are desirable for applied calibration training.

Gu, S., Kelly, B., and Xiu, D. (2021). "Autoencoder asset pricing models." In: *Journal of Econometrics* 222(1), pp. 429–450.

We propose a new latent factor conditional asset pricing model. Like Kelly, Pruitt, and Su (KPS, 2019), our model allows for latent factors and factor exposures that depend on covariates such as asset characteristics. But, unlike the linearity assumption of KPS, we model factor exposures as a flexible nonlinear function of covariates. Our model retrofits the workhorse unsupervised dimension reduction device from the machine learning literature - autoencoder neural networks - to incorporate information from covariates along with returns themselves. This delivers estimates of nonlinear conditional exposures and the associated latent factors. Furthermore, our machine learning framework imposes the economic restriction of no-arbitrage. Our autoencoder asset pricing model delivers out-of-sample pricing errors that are far smaller (and generally insignificant) compared to other leading factor models.

Gu, S., Kelly, B. T., and Xiu, D. (2020). "Empirical asset pricing via machine learning." In: *The Review of Financial Studies* 33 (5), pp. 2223–2273.

We synthesize the field of machine learning with the canonical problem of empirical asset pricing: Measuring asset risk premia. In the familiar empirical setting of cross section and time series stock return prediction, we perform a comparative analysis of methods in the machine learning repertoire, including generalize linear models, dimension reduction, boosted regression trees, random forests, and neural networks. At the broadest level, we find that machine learning offers an improved description of asset price behavior relative to traditional methods. Our implementation establishes a new standard for accuracy in measuring risk premia summarized by unprecedented high out-of-sample return prediction R2. We identify the best performing methods (trees and neural nets) and trace their predictive gains to allowance of nonlinear predictor interactions that are missed by other methods. Lastly, we find that all methods agree on the same small set of dominant predictive signals that includes variations on momentum, liquidity, and volatility. Improved risk premia measurement through machine learning can simplify the investigation into economic mechanisms of asset pricing and justifies its growing role in innovative financial technologies.

Haase, F. and Neuenkirch, M. (2021). "Forecasting Stock Market Recessions in the US: Predictive Modeling using Different Identification Approaches." In: *SSRN e-Print*.

The empirical literature of stock market predictability mainly suffers from model uncertainty and parameter instability. To meet this challenge, we propose a novel approach that combines the documented merits of diffusion indices, regime-switching models, and forecast combination to predict the dynamics in the S&P 500. First, we aggregate the weekly information of 115 popular macroeconomic and financial variables through an interaction of principal component analysis and shrinkage methods. Second, we estimate one-step Markov-switching models with time-varying transition probabilities using the diffusion indices as predictors. Third, we pool the forecasts in clusters to hedge against model risk and to evaluate the usefulness of different specifications. Our results show that we can adequately predict regime dynamics. Our forecasts provide a statistical improvement over several benchmarks and generate economic value by boosting returns, improving the certainty equivalent return, and reducing tail risk. Using the same approach for return forecasts, however, does not lead to a consistent outperformance of the historical average.

Habibnia, A. (2016). "Essays in high-dimensional nonlinear time series analysis." PhD thesis. London School of Economics and Political Science.
In this thesis, I study high-dimensional nonlinear time series analysis, and its applications in financial forecasting and identifying risk in highly interconnected financial networks. The first chapter is devoted to the testing for nonlinearity in financial time series. I present a tentative classification of the various linearity tests that have been proposed in the literature. Then I investigate nonlinear features of real financial series to determine if the data justify the use of nonlinear techniques, such as those inspired by machine learning theories. In Chapter 3 and 5, I develop forecasting strategies with a high-dimensional panel of predictors while considering nonlinear dynamics. Combining these two elements is a developing area of research. In the third chapter, I propose a nonlinear generalization of the statistical factor models. As a first step, factor estimation, I employ an auto-associative neural network to estimate nonlinear factors from predictors. In the second step, forecasting equation, I apply a nonlinear function -feedforward neural networkon estimated factors for prediction. I show that these features can go beyond covariance analysis and enhance forecast accuracy. I apply this approach to forecast equity returns, and show that capturing nonlinear dynamics between equities significantly improves the quality of forecasts over current univariate and multivariate factor models. In Chapter 5, I propose a high-dimensional learning based on a shrinkage estimation of a backpropagation algorithm for skip-layer neural networks. This thesis emphasizes that linear models can be represented as special cases of these two aforementioned models, which basically means that if there is no nonlinearity between series, the proposed models will reduce to a linear model. This thesis also includes a chapter (chapter 4, with Negar Kiyavash and Seyedjalal Etesami), which in this chapter, we propose a new approach for identifying and measuring systemic risk in financial networks by introducing a nonlinearly modified Granger-causality network based on directed information graphs. The suggested method allows for nonlinearity and has predictive power over future economic activity through a time-varying network of interconnections. We apply the method to the daily returns of U.S. financial Institutions including banks, brokers and insurance companiesto identifythe level of systemic risk inthe financial sector and the contribution of each financial institution.

Hammerschmid, R. and Lohre, H. (2018). "Regime Shifts and Stock Return Predictability." In: *International Review of Economics and Finance* 56, pp. 138–160.
Identifying economic regimes is useful in a world of time-varying risk premia. We apply regime switching models to common factors proxying for the macroeconomic regime and show that the ensuing regime factor is relevant in forecasting the equity risk premium. Moreover, the relevance of this regime factor is preserved in the presence of fundamental variables and technical indicators which are known to predict equity risk premia. Based on multiple predictive regressions and pooled forecasts, the macroeconomic regime factor is deemed complementary relative to the fundamental and technical information sets. Finally, these forecasts exhibit significant out-of-sample predictability that ultimately translates into considerable utility gains in a mean-variance portfolio strategy.

Hannadige, S. B., Gao, J., Silvapulle, M. J., and Silvapulle, P. (2021). "Forecasting a Nonstationary Time Series Using a Mixture of Stationary and Nonstationary Predictors." In: *SSRN e-Print*.
We develop a method for constructing prediction intervals for a nonstationary variable, such as GDP. The method uses a factor augmented regression [FAR] model. The predictors in the model includes a small number of factors generated to extract most of the information in a set of panel data on a large number of macroeconomic variables considered to be potential predictors. The novelty of this paper is that it provides a method and justification for a mixture of stationary and nonstationary factors as predictors in the FAR model; we refer to this as mixture-FAR method. This method is important because typically such a large set of panel data, for example the FRED-MD, is likely to contain a mixture of stationary and nonstationary variables. In our simulation study, we observed that the proposed mixture-FAR method performed better than its competitor that requires all the predictors to be nonstationary; the MSE of prediction was at least 33% lower for mixture-FAR. Using the data in FRED-QD for the US, we evaluated the aforementioned methods for forecasting the nonstationary variables, GDP and Industrial Production. We observed that the mixture-FAR method performed better than its competitors.

Harris, D., Martin, G. M., Perera, I., and Poskitt, D. S. (2019). "Construction and visualization of confidence sets for frequentist distributional forecasts." In: *Journal of Computational and Graphical Statistics* 28(2), pp. 92–104.

The focus of this paper is on the quantification of sampling variation in frequentist probabilistic forecasts. We propose a method of constructing confidence sets that respects the functional nature of the forecast distribution, and use animated graphics to visualize the impact of parameter uncertainty on the location, dispersion and shape of the distribution. The confidence sets are derived via the inversion of a Wald test and are asymptotically uniformly most accurate and, hence, optimal in this sense. A wide range of linear and non-linear time series models - encompassing long memory, state space and mixture specifications - is used to demonstrate the procedure, based on artificially generated data. An empirical example in which distributional forecasts of both financial returns and its stochastic volatility are produced is then used to illustrate the practical importance of accommodating sampling variation in the manner proposed.

Harvey, C. R. and Liu, Y. (2020). "Detecting Repeatable Performance." In: *SSRN e-Print*.

Past fund performance does a poor job of predicting future outcomes. The reason is noise. Using a random effects framework, we reduce the noise by pooling information from the cross-sectional alpha distribution to make density forecasts for each individual fund's alpha. In simulations, we show that our method generates parameter estimates that outperform alternative methods, both at the population and at the individual fund level. An out-of-sample forecasting exercise also shows that our method generates improved alpha forecasts.

Harvey, C. R., Liu, Y., and Saretto, A. (2020). "An Evaluation of Alternative Multiple Testing Methods for Finance Applications." In: *The Review of Asset Pricing Studies* 10(2), pp. 199–248.

In almost every area of empirical finance, researchers confront multiple tests. One high-profile example is the identification of outperforming investment managers, many of whom beat their benchmarks purely by luck. Multiple testing methods are designed to control for luck. Factor selection is another glaring case in which multiple tests are performed, but numerous other applications do not receive as much attention. One important example is a simple regression model testing five variables. In this case, because five variables are tried, a t-statistic of 2.0 is not enough to establish significance. Our paper provides a guide to various multiple testing methods and details a number of applications. We provide simulation evidence on the relative performance of different methods across a variety of testing environments. The goal of our paper is to provide a menu that researchers can choose from to improve inference in financial economics.

Harvey, D. I., Leybourne, S. J., Sollis, R., and Taylor, A. M. R. (2021). "Real-Time Detection of Regimes of Predictability in the U.S. Equity Premium." In: *Journal of Applied Econometrics* 36, pp. 45–70.

We propose new real-time monitoring procedures for the emergence of end-of-sample predictive regimes using sequential implementations of standard (heteroskedasticity-robust) regression t-statistics for predictability applied over relatively short time periods. The procedures we develop can also be used for detecting historical regimes of temporary predictability. Our proposed methods are robust to both the degree of persistence and endogeneity of the regressors in the predictive regression and to certain forms of heteroskedasticity in the shocks. We discuss how the monitoring procedures can be designed such that their false positive rate can be set by the practitioner at the start of the monitoring period using detection rules based on information obtained from the data in a training period. We use these new monitoring procedures to investigate the presence of regime changes in the predictability of the US equity premium at the 1-month horizon by traditional macroeconomic and financial variables, and by binary technical analysis indicators. Our results suggest that the 1-month-ahead equity premium has temporarily been predictable, displaying so-called ”pockets of predictability," and that these episodes of predictability could have been detected in real time by practitioners using our proposed methodology.

Hassler, U. and Pohle, M.-O. (2021). "Forecasting under Long Memory." In: *Journal of Financial Econometrics*.

Motivated by the mixed evidence in the literature on forecasting long memory processes, we show that methods based on fractional integration are superior to alternatives not accounting for long memory by simulations and applications to classical long memory time series from macroeconomics and finance. Furthermore, we analyze the optimal implementation of these methods, among others comparing parametric and local and global semi-parametric estimators of the long memory parameter, providing asymptotic theory on different mean estimators and assessing the use of a fixed long memory parameter to overcome the inherent difficulties of its estimation.

Hauzenberger, N., Huber, F., and Klieber, K. (2021). "Real-time Inflation Forecasting Using Non-linear Dimension Reduction Techniques." In: *arXiv e-Print*.

In this paper, we assess whether using non-linear dimension reduction techniques pays off for forecasting inflation in real-time. Several recent methods from the machine learning literature are adopted to map a large dimensional dataset into a lower dimensional set of latent factors. We model the relationship between inflation and the latent factors using constant and time-varying parameter (TVP) regressions with shrinkage priors. Our models are then used to forecast monthly US inflation in real-time. The results suggest that sophisticated dimension reduction methods yield inflation forecasts that are highly competitive to linear approaches based on principal components. Among the techniques considered, the Autoencoder and squared principal components yield factors that have high predictive power for one-month- and one-quarter-ahead inflation. Zooming into model performance over time reveals that controlling for non-linear relations in the data is of particular importance during recessionary episodes of the business cycle or the current COVID-19 pandemic.

He, S. and Gu, S. (2022). "Multi-modal Attention Network for Stock Movements Prediction." In: *arXiv e-Print*.
Stock prices move as piece-wise trending fluctuation rather than a purely random walk. Traditionally, the prediction of future stock movements is based on the historical trading record. Nowadays, with the development of social media, many active participants in the market choose to publicize their strategies, which provides a window to glimpse over the whole market's attitude towards future movements by extracting the semantics behind social media. However, social media contains conflicting information and cannot replace historical records completely. In this work, we propose a multi-modality attention network to reduce conflicts and integrate semantic and numeric features to predict future stock movements comprehensively. Specifically, we first extract semantic information from social media and estimate their credibility based on posters' identity and public reputation. Then we incorporate the semantic from online posts and numeric features from historical records to make the trading strategy. Experimental results show that our approach outperforms previous methods by a significant margin in both prediction accuracy (61.20%) and trading profits (9.13%). It demonstrates that our method improves the performance of stock movements prediction and informs future research on multi-modality fusion towards stock prediction.

Herzen, J., Lassig, F., Piazzetta, S. G., Neuer, T., Tafti, L., Raille, G., Pottelbergh, T. V., Pasieka, M., Skrodzki, A., Huguenin, N., Dumonal, M., Koscisz, J., Bader, D., Gusset, F., Benheddi, M., Williamson, C., Kosinski, M., Petrik, M., and Grosch, G. (2022). "Darts: User-Friendly Modern Machine Learning for Time Series." In: *Journal of Machine Learning Research* 23(124), pp. 1–6.
We present Darts, a Python machine learning library for time series, with a focus on forecasting. Darts offers a variety of models, from classics such as ARIMA to state-of-the-art deep neural networks. The emphasis of the library is on offering modern machine learning functionalities, such as supporting multidimensional series, meta-learning on multiple series, training on large datasets, incorporating external data, ensembling models, and providing a rich support for probabilistic forecasting. At the same time, great care goes into the API design to make it user-friendly and easy to use. For instance, all models can be used using fit()/predict(), similar to scikit-learn.

Hewamalage, H., Bergmeir, C., and Bandara, K. (2020). "Global Models for Time Series Forecasting: A Simulation Study." In: *arXiv e-Print*.
In the current context of Big Data, the nature of many forecasting problems has changed from predicting isolated time series to predicting many time series from similar sources. This has opened up the opportunity to develop competitive global forecasting models that simultaneously learn from many time series. But, it still remains unclear when global forecasting models can outperform the univariate benchmarks, especially along the dimensions of the homogeneity/heterogeneity of series, the complexity of patterns in the series, the complexity of forecasting models, and the lengths/number of series. Our study attempts to address this problem through investigating the effect from these factors, by simulating a number of datasets that have controllable time series characteristics. Specifically, we simulate time series from simple data generating processes (DGP), such as Auto Regressive (AR) and Seasonal AR, to complex DGPs, such as Chaotic Logistic Map, Self-Exciting Threshold Auto-Regressive, and Mackey-Glass Equations. The data heterogeneity is introduced by mixing time series generated from several DGPs into a single dataset. The lengths and the number of series in the dataset are varied in different scenarios. We perform experiments on these datasets using global forecasting models including Recurrent Neural Networks (RNN), Feed-Forward Neural Networks, Pooled Regression (PR) models and Light Gradient Boosting Models (LGBM), and compare their performance against standard statistical univariate forecasting techniques. Our experiments demonstrate that when trained as global forecasting models, techniques such as RNNs and LGBMs, which have complex non-linear modelling capabilities, are competitive methods in general under challenging forecasting scenarios such as series having short lengths, datasets with heterogeneous series and having minimal prior knowledge of the patterns of the series.

Hewamalage, H., Bergmeir, C., and Bandara, K. (2021). "Recurrent Neural Networks for Time Series Forecasting: Current status and future directions." In: *International Journal of Forecasting* 37(1), pp. 388–427.
Recurrent Neural Networks (RNNs) have become competitive forecasting methods, as most notably shown in the winning method of the recent M4 competition. However, established statistical models such as exponential

smoothing (ETS) and the autoregressive integrated moving average (ARIMA) gain their popularity not only from their high accuracy, but also because they are suitable for non-expert users in that they are robust, efficient, and automatic. In these areas, RNNs have still a long way to go. We present an extensive empirical study and an open-source software framework of existing RNN architectures for forecasting, and we develop guidelines and best practices for their use. For example, we conclude that RNNs are capable of modelling seasonality directly if the series in the dataset possess homogeneous seasonal patterns; otherwise, we recommend a deseasonalisation step. Comparisons against ETS and ARIMA demonstrate that (semi-) automatic RNN models are not silver bullets, but they are nevertheless competitive alternatives in many situations.

Ho, T.-W. and Lin, Y.-c. (2021). "Training by Rolling: Machine Learning and Stock Returns Forecasting." In: *SSRN e-Print*.

Stock returns predictability has been a long-standing puzzle in literature of financial economics, and recent developments in prediction technology have populated the use of machine learning techniques, which motivate our study to whether the stock returns predictability (Chen, Roll & Ross, 1986) can be improved by machine learning. We contribute to the literature by methodologically proposing the training-by-rolling framework: First, the K-fold cross validation is embedded into a rolling framework to search for the best hyperparameters and model settings, that is, each roll is trained by K-fold cross validation. Second, to obtain the model-specific rolling prediction, we apply model estimates of rolling validations to forecast the testing data. Third, to obtain the training prediction, we combine the predictions of the test data as K-fold CV does, and we then finally average across all models to show that models average further improve prediction. Our study confirm that machine learning methods outperform the standard ARMA-based time series methods in predicting the equity returns.

Hofmarcher, P. and Grun, B. (2020). "Bayesian Model Averaging." In: *Macroeconomic Forecasting in the Era of Big Data*. Springer International Publishing, pp. 359–388.

Bayesian model averaging (BMA) is a statistical method to rigorously take model uncertainty into account. This chapter gives a coherent overview on the statistical foundations and methods of BMA and its usefulness for forecasting, but also for the identification of robust determinants. The focus is given on economic applications. We describe the BMA framework in the context of linear models. Different model and parameter priors are discussed in detail and suitable inference methods and posterior analysis tools are presented. We illustrate the use of the presented BMA framework to study potential drivers of box office revenues and to forecast these revenues. The available data set does not only contain traditional variables such as budget or genre categorization, but also variables derived from social media content which capture sentiment and volume of Twitter messages. We discuss the impact of different model specifications and describe how results are obtained using the open-source package BMS available for the R environment for statistical computing and graphics.

Hoga, Y. and Dimitriadis, T. (2021). "On Testing Equal Conditional Predictive Ability Under Measurement Error." In: *arXiv e-Print*.

Loss functions are widely used to compare several competing forecasts. However, forecast comparisons are often based on mismeasured proxy variables for the true target. We introduce the concept of exact robustness to measurement error for loss functions and fully characterize this class of loss functions as the Bregman class. For such exactly robust loss functions, forecast loss differences are on average unaffected by the use of proxy variables and, thus, inference on conditional predictive ability can be carried out as usual. Moreover, we show that more precise proxies give predictive ability tests higher power in discriminating between competing forecasts. Simulations illustrate the different behavior of exactly robust and non-robust loss functions. An empirical application to US GDP growth rates demonstrates that it is easier to discriminate between forecasts issued at different horizons if a better proxy for GDP growth is used.

Hollyman, R., Petropoulos, F., and Tipping, M. E. (2021). "Understanding forecast reconciliation." In: *European Journal of Operational Research* 294(1), pp. 149–160.

A series of recent papers introduce the concept of Forecast Reconciliation, a process by which independently generated forecasts of a collection of linearly related time series are reconciled via the introduction of accounting aggregations that naturally apply to the data. Aside from its clear presentational and operational virtues, the reconciliation approach generally improves the accuracy of the combined forecasts. In this paper, we examine the mechanisms by which this improvement is generated by re-formulating the reconciliation problem as a combination of direct forecasts of each time series with additional indirect forecasts derived from the linear constraints. Our work establishes a direct link between the nascent Forecast Reconciliation literature and the extensive work on Forecast Combination. In the original hierarchical setting, our approach clarifies for the first time how unbiased forecasts for the entire collection can be generated from base forecasts made at any level of the hierarchy, and we illustrate more generally how simple robust combined forecasts can be generated in any multivariate setting subject to linear constraints. In an empirical example, we show that simple combinations of such forecasts generate significant improvements in forecast accuracy where it matters most: where noise levels are highest and the forecasting task is at its most challenging.

Holzmann, H. and Klar, B. (2021). "Using Proxies to Improve Forecast Evaluation." In: *arXiv e-Print*.
Comparative evaluation of forecasts of statistical functionals relies on comparing averaged losses of competing forecasts after the realization of the quantity Y, on which the functional is based, has been observed. Motivated by high-frequency finance, in this paper we investigate how proxies $\tilde{Y}$ for Y - say volatility proxies - which are observed together with Y can be utilized to improve forecast comparisons. We extend previous results on robustness of loss functions for the mean to general moments and ratios of moments, and show in terms of the variance of differences of losses that using proxies will increase the power in comparative forecast tests. These results apply both to testing conditional as well as unconditional dominance. Finally, we illustrate the theoretical results for simulated high-frequency data.

Hounyo, U. and Lahiri, K. (2022). "Estimating the variance of a combined forecast: Bootstrap-based approach." In: *Journal of Econometrics*.
This paper considers bootstrap inference in model averaging for predictive regressions. We first show that the standard pairwise bootstrap is not valid in the context of model averaging. This common bootstrap approach induces a bias-related term in the bootstrap variance of averaging estimators. We then propose and justify a fixed-design residual-based bootstrap resampling approach for model averaging. In a local asymptotic framework, we show the validity of the bootstrap in estimating the variance of a combined forecast and the asymptotic covariance matrix of a combined parameter vector with fixed weights. Our proposed method preserves nonparametrically the cross-sectional dependence between different models and the time series dependence in the errors simultaneously. The finite sample performance of these methods is assessed via Monte Carlo simulations. We illustrate our approach using an empirical study of the Taylor rule equation with 24 alternative specifications.

Hsiao, C. and Wan, S. K. (2014). "Is there an optimal forecast combination?" In: *Journal of Econometrics* 178, pp. 294–309.
We consider several geometric approaches for combining forecasts in large samples-a simple eigenvector approach, a mean corrected eigenvector and trimmed eigenvector approach. We give conditions where geometric approach yields identical result as the regression approach. We also consider a mean and scale corrected simple average of all predictive models for finite sample and give conditions where simple average is an optimal combination. Monte Carlos are conducted to compare the finite sample performance of these and some popular forecast combination and information combination methods and to shed light on the issues of ”forecast combination”vs ”information combination”. We also try to shed light on whether there exists an optimal forecast combination method by comparing various forecast combination methods to predict US real output growth rate and excess equity premium.

Hull, B. and Qiao, X. (2017). "A Practitioner's Defense of Return Predictability." In: *The Journal of Portfolio Management* 43(3), pp. 60–76.
Revisiting the issue of return predictability, the authors show there is substantial predictive power in combining forecasting variables. Applying correlation screening to combine 20 variables that have been proposed in the return predictability literature, the authors demonstrate forecasting power at a six-month horizon. They illustrate the economic significance of return predictability through a walk-forward simulation, which takes positions in the SandP 500 ETF Trust (SPY) proportional to the model forecast equity risk premium. The simulated strategy yields annual returns more than twice that of the buy-and-hold strategy, with a Sharpe ratio four times as large. To eliminate look-ahead bias, the authors perform additional simulations, while including variables only as they are discovered in the literature. Results show similar annual returns and Sharpe ratios. Although a market-timing strategy outperforms the market, the authors maintain that it is difficult to implement.

Hunt, I. (2022). "In-sample tests of predictability are superior to pseudo-out-of-sample tests, even when data mining." In: *International Journal of Forecasting*.
This paper analyses straightforward Bonferroni adjustments to critical values of in-sample tests of predictability, when data mining is used to search across models. Unlike conventional pseudo-out-of-sample tests, these in-sample tests have stable family-wise error rates (FWERs) when searching for models that predict well. Furthermore, when data mining, these in-sample tests have more power than pseudo-out-of-sample tests for identifying true predictability.

Hyndman, R. J. (2020). "Tidy Time Series and Forecasting in R." In: *RStudio conf2020*.
On day 1, we looked at the tsibble, lubridate and feasts packages (along with the tidyverse of course). We introduced the tsibble data structure for flexibly managing collections of related time series, and explored how to do data wrangling, data visualizations and exploratory data analysis, along with some feature-based methods to explore time series data in high dimensions. Day 2 was about forecasting using the fable package. We looked at several well-known time series forecasting models and how they are automated in the fable package. We also discussed ensemble forecasts. Finally, we looked at forecast reconciliation, allowing millions of time series to be forecast in a relatively short time while accounting for constraints on how the series are related.

Hyndman, R. J. and Athanasopoulos, G. (2020). *Forecasting: Principles and Practice (Third Edition)*. OTexts. 380 pp.
Forecasting is required in many situations. Deciding whether to build another power generation plant in the next five years requires forecasts of future demand. Scheduling staff in a call centre next week requires forecasts of call volumes. Stocking an inventory requires forecasts of stock requirements. Telecommunication routing requires traffic forecasts a few minutes ahead. Whatever the circumstances or time horizons involved, forecasting is an important aid in effective and efficient planning. This textbook provides a comprehensive introduction to forecasting methods and presents enough information about each method for readers to use them sensibly. Examples use R with many data sets taken from the authors' own consulting experience. In this second edition, all chapters have been updated to cover the latest research and forecasting methods. Three new chapters have been added on dynamic regression forecasting, hierarchical forecasting and practical forecasting issues. The latest version of the book is freely available online at `http://OTexts.com/fpp3`.

Ilic, I., Gorgulu, B., and Cevik, M. (2020). "Augmented Out-of-Sample Comparison Method for Time Series Forecasting Techniques." In: *Advances in Artificial Intelligence*. Springer International Publishing, pp. 302–308.
Time series data consists of high dimensional sets of observations with strong spatio-temporal relations. Accordingly, conventional methods for comparing different regression methods, such as random train-test splits, do not sufficiently evaluate time series forecasting tasks. In this work, we introduce a robust technique for out-of-sample forecasting that takes the spatio-temporal nature of time series into account. We compare well-known auto-regressive integrated moving average (ARIMA) models with recurrent neural network (RNN) based models using Turkish electricity data. We observe that RNN-based models outperform ARIMA models. Moreover, as the length of forecast interval increases, the performance gap widens between these two approaches.

Ilmanen, A., Chandra, S., and McQuinn, N. (2020). "Demystifying illiquid assets: expected returns for private equity." In: *The Journal of Alternative Investments* 22(3), pp. 8–22.
The growing interest in private equity means that allocators must carefully evaluate its risk and return. The challenge is that modeling private equity is not straightforward, due to a lack of good quality data and artificially smooth returns. We try to demystify the subject, considering theoretical arguments, historical average returns, and a forward-looking analysis. For institutional investors trying to calibrate their asset allocation decisions for private equity, we lay out a framework for expected returns, albeit one hampered by data limitations, that is based on a discounted cash-flow framework similar to what we use for public stocks and bonds. In particular, we attempt to assess private equity's realized and estimated expected return edges over lower-cost public equity counterparts. Our estimates display a decreasing trend over time, which does not seem to have slowed the institutional demand for private equity. We conjecture that this is due to investors' preference for the return-smoothing properties of illiquid assets in general.

Inoue, A., Jin, L., and Rossi, B. (2017). "Rolling window selection for out-of-sample forecasting with time-varying parameters." In: *Journal of Econometrics* 196(1), pp. 55–67.
There is strong evidence of structural changes in macroeconomic time series, and the forecasting performance is often sensitive to the choice of estimation window size. This paper develops a method for selecting the window size for forecasting. Our proposed method is to choose the optimal size that minimizes the forecaster's quadratic loss function, and we prove the asymptotic validity of our approach. Our Monte Carlo experiments show that our method performs well under various types of structural changes. When applied to forecasting US real output growth and inflation, the proposed method tends to improve upon conventional methods, especially for output growth.

Iworiso, J. and Vrontos, S. (2021). "On the Predictability of the Equity Premium Using Deep Learning Techniques." In: *The Journal of Financial Data Science* 3(1), pp. 74–92.
Deep learning is drawing keen attention in contemporary financial research. In this article, the authors investigate the statistical predictive power and economic significance of financial stock market data by using deep learning

techniques. In particular, the authors use the equity premium as the response variable and financial variables as predictors. The deep learning techniques used in this study provide useful evidence of statistical predictability and economic significance. Considering the statistical predictive performance of the deep learning models, H2O deep learning (H2ODL) gives the smallest mean-squared forecast error (MSFE), with the corresponding highest cumulative return (CR) and Sharpe ratio (SR) in each of the out-of-sample periods. Specifically, the H2ODL with Rectifier used as the activation function outperformed the other models in this article. In the fusion results, the SAE-with-H2O using the Maxout activation function yields the smallest MSFE with the corresponding highest CR and SR in all of the out-of-sample periods. It is worth noting that the higher the CR, the higher the SR and the lower the MSFE, which concords with a rule of thumb. Overall, the empirical analysis in this study revealed that the SAE-with-H2O using the Maxout activation function produced the best statistically predictive and economically significant results with robustness across all out-of-sample periods.

Jaganathan, S. and Prakash, P. K. S. (2020). "A combination-based forecasting method for the M4-competition." In: *International Journal of Forecasting* 36(1), pp. 98–104.

Several researchers (Armstrong, 2001; Clemen, 1989; Makridakis and Winkler, 1983) have shown empirically that combination-based forecasting methods are very effective in real world settings. This paper discusses a combination-based forecasting approach that was used successfully in the M4 competition. The proposed approach was evaluated on a set of 100K time series across multiple domain areas with varied frequencies. The point forecasts submitted finished fourth based on the overall weighted average (OWA) error measure and second based on the symmetric mean absolute percent error (sMAPE).

Janssen, R. V. (2019). "Multi-horizon comparison of multivariate inflation forecasting." MA thesis. Erasmus School of Economics.

This paper applies the multi-horizon comparison methodology from Quaedvlieg (2019) to assess the forecasting performance of direct and iterative multivariate inflation forecasts, with both highand low lagorders. We use variousmacroeconomic indicators ina GETS restricted estimation to forecastUSinflation and show that high orderVARs on average prefer iterative forecasts, while low order VARs on average prefer the direct forecasts. Finally, we provide evidence that the best high order multivariate forecasts outperform the best low order multivariate forecastson every individual horizon(uniform superior predictive abilities). Thisimpliesthat in this setting, inflation forecasts are most accurately forecasted with a high orderVAR using an iterativ eapproach.

Januschowski, T., Wang, Y., Torkkola, K., Erkkila, T., Hasson, H., and Gasthaus, J. (2022). "Forecasting with trees." In: *International Journal of Forecasting*.

The prevalence of approaches based on gradient boosted trees among the top contestants in the M5 competition is potentially the most eye-catching result. Tree-based methods out-shone other solutions, in particular deep learning-based solutions. The winners in both tracks of the M5 competition heavily relied on them. This prevalence is even more remarkable given the dominance of other methods in the literature and the M4 competition. This article tries to explain why tree-based methods were so widely used in the M5 competition. We see possibilities for future improvements of tree-based models and then distill some learnings for other approaches, including but not limited to neural networks.

Jegadeesh, N., Noh, J., Pukthuanthong, K., Roll, R., and Wang, J. (2019). "Empirical tests of asset pricing models with individual assets: Resolving the errors-in-variables bias in risk premium estimation." In: *Journal of Financial Economics* 113(2), pp. 273–298.

To attenuate an inherent errors-in-variables bias, portfolios are widely employed to test asset pricing models; but portfolios might mask relevant risk- or return-related features of individual stocks. We propose an instrumental variables approach that allows the use of individual stocks as test assets, yet delivers consistent estimates of ex post risk premiums. This estimator also yields well-specified tests in small samples. The market risk premium under the capital asset pricing model (CAPM) and the liquidity-adjusted CAPM, premiums on risk factors under the Fama-French three- and five-factor models, and the Hou et al. (2015) four-factor model are all insignificant after controlling for asset characteristics.

Jin, S., Corradi, V., and Swanson, N. R. (2017). "Robust Forecast Comparison." In: *Econometric Theory* 33(06), pp. 1306–1351.

Forecast accuracy is typically measured in terms of a given loss function. However, as a consequence of the use of misspecified models in multiple model comparisons, relative forecast rankings are loss function dependent. In order to address this issue, a novel criterion for forecast evaluation that utilizes the entire distribution of forecast errors is introduced. In particular, we introduce the concepts of general-loss (GL) forecast superiority and convex-loss (CL) forecast superiority; and we develop tests for GL (CL) superiority that are based on an out-of-sample generalization of the tests introduced by Linton, Maasoumi, and Whang (2005,Review of Economic Studies 72, 735-765). Our test statistics are characterized by nonstandard limiting distributions, under the null, necessitating the use of resampling procedures to obtain critical values. Additionally, the tests are consistent and have nontrivial local power, under a sequence of local alternatives. The above theory is developed for the stationary case, as well as for the case of heterogeneity that is induced by distributional change over time. Monte Carlo simulations suggest that the tests perform reasonably well in finite samples, and an application in which we examine exchange rate data indicates that our tests can help identify superior forecasting models, regardless of loss function.

Jordan, A., Kruger, F., and Lerch, S. (2019). "Evaluating probabilistic forecasts with scoringRules." In: *Journal of Statistical Software* 90(12).
Probabilistic forecasts in the form of probability distributions over future events have become popular in several fields including meteorology, hydrology, economics, and demography. In typical applications, many alternative statistical models and data sources can be used to produce probabilistic forecasts. Hence, evaluating and selecting among competing methods is an important task. The scoringRules package for R provides functionality for comparative evaluation of probabilistic models based on proper scoring rules, covering a wide range of situations in applied work. This paper discusses implementation and usage details, presents case studies from meteorology and economics, and points to the relevant background literature.

Joshi, S. (2019). "Time Series Analysis and Forecasting of the US Housing Starts using Econometric and Machine Learning Model." In: *arXiv e-Print*.
In this research paper, I have performed time series analysis and forecasted the monthly value of housing starts for the year 2019 using several econometric methods - ARIMA(X), VARX, (G)ARCH and machine learning algorithms - artificial neural networks, ridge regression, K-Nearest Neighbors, and support vector regression, and created an ensemble model. The ensemble model stacks the predictions from various individual models, and gives a weighted average of all predictions. The analyses suggest that the ensemble model has performed the best among all the models as the prediction errors are the lowest, while the econometric models have higher error rates.

Kalfa, S. Y. and Marquez, J. (2021). "Forecasting FOMC Forecasts." In: *Econometrics* 9(3), p. 34.
(Hendry 1980, p. 403) The three golden rules of econometrics are ”test, test, and test”. The current paper applies that approach to model the forecasts of the Federal Open Market Committee over 1992-2019 and to forecast those forecasts themselves. Monetary policy is forward-looking, and as part of the FOMC's effort toward transparency, the FOMC publishes its (forward-looking) economic projections. The overall views on the economy of the FOMC participants-as characterized by the median of their projections for inflation, unemployment, and the Fed's policy rate-are themselves predictable by information publicly available at the time of the FOMC's meeting. Their projections also communicate systematic behavior on the part of the FOMC's participants.

Kamarthi, H., Kong, L., Rodriguez, A., Zhang, C., and Prakash, B. A. (2021). "CAMul: Calibrated and Accurate Multi-view Time-Series Forecasting." In: *arXiv e-Print*.
Probabilistic time-series forecasting enables reliable decision making across many domains. Most forecasting problems have diverse sources of data containing multiple modalities and structures. Leveraging information as well as uncertainty from these data sources for well-calibrated and accurate forecasts is an important challenging problem. Most previous work on multi-modal learning and forecasting simply aggregate intermediate representations from each data view by simple methods of summation or concatenation and do not explicitly model uncertainty for each data-view. We propose a general probabilistic multi-view forecasting framework CAMul, that can learn representations and uncertainty from diverse data sources. It integrates the knowledge and uncertainty from each data view in a dynamic context-specific manner assigning more importance to useful views to model a well-calibrated forecast distribution. We use CAMul for multiple domains with varied sources and modalities and show that CAMul outperforms other state-of-art probabilistic forecasting models by over 25% in accuracy and calibration.

Kang, T.-H., Sharma, A., and Marshall, L. (2021a). "Assessing Goodness of Fit for Verifying Probabilistic Forecasts." In: *Forecasting* 3(4), pp. 763–773.
The verification of probabilistic forecasts in hydro-climatology is integral to their development, use, and adoption. We propose here a means of utilizing goodness of fit measures for verifying the reliability of probabilistic forecasts. The difficulty in measuring the goodness of fit for a probabilistic prediction or forecast is that predicted probability distributions for a target variable are not stationary in time, meaning one observation alone exists to quantify goodness of fit for each prediction issued. Therefore, we suggest an additional dissociation

that can dissociate target information from the other time variant part-the target to be verified in this study is the alignment of observations to the predicted probability distribution. For this dissociation, the probability integral transformation is used. To measure the goodness of fit for the predicted probability distributions, this study uses the root mean squared deviation metric. If the observations after the dissociation can be assumed to be independent, the mean square deviation metric becomes a chi-square test statistic, which enables statistically testing the hypothesis regarding whether the observations are from the same population as the predicted probability distributions. An illustration of our proposed rationale is provided using the multi-model ensemble prediction for El Nino-Southern Oscillation.

Kang, Y., Spiliotis, E., Petropoulos, F., Athiniotis, N., Li, F., and Assimakopoulos, V. (2021b). "Déjà vu: A data-centric forecasting approach through time series cross-similarity." In: *Journal of Business Research*.
Accurate forecasts are vital for supporting the decisions of modern companies. Forecasters typically select the most appropriate statistical model for each time series. However, statistical models usually presume some data generation process while making strong assumptions about the errors. In this paper, we present a novel data-centric approach – 'forecasting with cross-similarity', which tackles model uncertainty in a model-free manner. Existing similarity-based methods focus on identifying similar patterns within the series, i.e., 'self-similarity'. In contrast, we propose searching for similar patterns from a reference set, i.e., 'cross-similarity'. Instead of extrapolating, the future paths of the similar series are aggregated to obtain the forecasts of the target series. Building on the cross-learning concept, our approach allows the application of similarity-based forecasting on series with limited lengths. We evaluate the approach using a rich collection of real data and show that it yields competitive accuracy in both points forecasts and prediction intervals.

Karathanasopoulos, A., Mitra, S., Skindilias, K., and Lo, C. C. (2017). "Modelling and Trading the English and German Stock Markets with Novelty Optimization Techniques." In: *Journal of Forecasting* 36(8) (8), pp. 974–988.
The motivation for this paper was the introduction of novel short-term models to trade the FTSE 100 and DAX 30 exchange-traded funds (ETF) indices. There are major contributions in this paper which include the introduction of an input selection criterion when utilizing an expansive universe of inputs, a hybrid combination of partial swarm optimizer (PSO) with radial basis function (RBF) neural networks, the application of a PSO algorithm to a traditional autoregressive moving model (ARMA), the application of a PSO algorithm to a higher-order neural network and, finally, the introduction of a multi-objective algorithm to optimize statistical and trading performance when trading an index. All the machine learning-based methodologies and the conventional models are adapted and optimized to model the index. A PSO algorithm is used to optimize the weights in a traditional RBF neural network, in a higher-order neural network (HONN) and the AR and MA terms of an ARMA model. In terms of checking the statistical and empirical accuracy of the novel models, we benchmark them with a traditional HONN, with an ARMA, with a moving average convergence/divergence model (MACD) and with a naive strategy. More specifically, the trading and statistical performance of all models is investigated in a forecast simulation of the FTSE 100 and DAX 30 ETF time series over the period January 2004 to December 2015 using the last 3 years for out-of-sample testing. Finally, the empirical and statistical results indicate that the PSO-RBF model outperforms all other examined models in terms of trading accuracy and profitability, even with mixed inputs and with only autoregressive inputs.

Karolyi, A. and Van Nieuwerburgh, S. (2020). "New Methods for the Cross-Section of Returns." In: *The Review of Financial Studies* 33(5), pp. 1879–1890.
The cross-section and time series of stock returns contains a wealth of information about the stochastic discount factor (SDF), the object that links cash flows to prices. A large empirical literature has uncovered many candidate factors many more than seem plausible to summarize the SDF. This special volume of the Review of Financial Studies presents recent advances in extracting information from both the cross-section and the time series, in dealing with issues of replication and false discoveries, and in applying innovative machine-learning techniques to identify the most relevant asset pricing factors. Our editorial summarizes what we learn and offers a few suggestions to guide future work in this exciting new era of big data and empirical asset pricing.

Kelly, B. T., Pruitt, S., and Su, Y. (2019). "Characteristics are covariances: A unified model of risk and return." In: *Journal of Financial Economics* 134(3), pp. 501–524.
We propose a new modeling approach for the cross section of returns. Our method, Instrumented Principal Component Analysis (IPCA), allows for latent factors and time-varying loadings by introducing observable characteristics that instrument for the unobservable dynamic loadings. If the characteristics/expected return relationship is driven by compensation for exposure to latent risk factors, IPCA will identify the corresponding

latent factors. If no such factors exist, IPCA infers that the characteristic effect is compensation without risk and allocates it to an intercept. Studying returns and characteristics at the stock-level, we find that five IPCA factors explain the cross section of average returns significantly more accurately than existing factor models and produce characteristic-associated anomaly intercepts that are small and statistically insignificant. Furthermore, among a large collection of characteristics explored in the literature, only ten are statistically significant at the 1% level in the IPCA specification and are responsible for nearly 100% of the model accuracy.

Kiefer, D., Bauer, M., and Grimm, F. (2021). "Univariate Time Series Forecasting: Machine Learning Prediction of the Best Suitable Forecast Model Based on Time Series Characteristics." In: *Human Centred Intelligent Systems*. Springer Singapore, pp. 152–162.
Forecasting demand is challenging. Various products exhibit different demand patterns. While demand may be constant and regular for one product, it may be sporadic for another, as well as when demand occurs, it may fluctuate significantly. Forecasting errors are costly and result in obsolete inventory or unsatisfied demand. Methods from statistics, machine learning, and deep learning have been used to predict such demand patterns. Nevertheless, it is not clear for what demand pattern, which algorithm would achieve the best forecast. Therefore, even today a large number of models are used to forecast on a test period. The model with the best result on the test period is used for the actual forecast. This approach is computationally and time intensive and, in most cases, uneconomical. In our paper we show the possibility to use a machine learning classification algorithm, which predicts the best possible model based on the characteristics of a time series. The approach was developed and evaluated on a dataset from a B2B-technical-retailer. The machine learning classification algorithm achieves a mean ROC-AUC of 89%, which emphasizes the skill of the model.

Klingberg Malmer, O. and Pettersson, G. (2020). "Tidying up the factor zoo: Using machine learning to find sparse factor models that predict asset returns." MA thesis. University of Goteborg.
There exist over 300 firm characteristics that provide significant information about average asset return. John Cochrane refers to this as a "factor zoo" and challenges researchers to find the independent characteristics which can explain average return. That is, to find the unsubsumed and non-nested firm characteristics that are highly predictive of asset return. In this thesis we act on the posed challenge by using a data driven approach. We apply two machine learning methods to create sparse factor models composed by a small set of these characteristics. The two methods are one unsupervised learning method, the Principal Component Analysis, and one supervised learning method, the LASSO regression. The study is done using the S&P 500 index constituents and 54 firm characteristics over the time period 2009-07-01 to 2019-07-01. The performance of the factor models is in this study measured using out-of-sample measurements. Using established methods of post-LASSO regression and new developed techniques for variable selection based on PCA, we generate four new factor models. The latter mentioned variable selection method based on PCA is, to our knowledge, an original contribution of this thesis. The generated factor models are compared against the Fama French factors in the out-of-sample test and are shown to all outperform. The best performer is a LASSO generated factor model containing 6 factors. By analysing the results we find that momentum factors, such as price relative to 52-week-high-price, are highly predictive of return and are commonly selected factors, which confirms the results of previous responses to the same challenge.

Kosman, E. and Castro, D. D. (2021). "Vision-Guided Forecasting – Visual Context for Multi-Horizon Time Series Forecasting." In: *arXiv e-Print*.
Autonomous driving gained huge traction in recent years, due to its potential to change the way we commute. Much effort has been put into trying to estimate the state of a vehicle. Meanwhile, learning to forecast the state of a vehicle ahead introduces new capabilities, such as predicting dangerous situations. Moreover, forecasting brings new supervision opportunities by learning to predict richer a context, expressed by multiple horizons. Intuitively, a video stream originated from a front-facing camera is necessary because it encodes information about the upcoming road. Besides, historical traces of the vehicle's states give more context. In this paper, we tackle multi-horizon forecasting of vehicle states by fusing the two modalities. We design and experiment with 3 end-to-end architectures that exploit 3D convolutions for visual features extraction and 1D convolutions for features extraction from speed and steering angle traces. To demonstrate the effectiveness of our method, we perform extensive experiments on two publicly available real-world datasets, Comma2k19 and the Udacity challenge. We show that we are able to forecast a vehicle's state to various horizons, while outperforming the current state-of-the-art results on the related task of driving state estimation. We examine the contribution of vision features, and find that a model fed with vision features achieves an error that is 56.6% and 66.9% of the error of a model that doesn't use those features, on the Udacity and Comma2k19 datasets respectively.

Koutsandreas, D., Spiliotis, E., Petropoulos, F., and Assimakopoulos, V. (2021). "On the selection of forecasting accuracy measures." In: *Journal of the Operational Research Society*.

A lot of controversy exists around the choice of the most appropriate error measure for assessing the performance of forecasting methods. While statisticians argue for the use of measures with good statistical properties, practitioners prefer measures that are easy to communicate and understand. Moreover, researchers argue that the loss-function for parameterizing a model should be aligned with how the post-performance measurement is made. In this paper we ask: Does it matter? Will the relative ranking of the forecasting methods change significantly if we choose one measure over another? Will a mismatch of the in-sample loss-function and the out-of-sample performance measure decrease the performance of the forecasting models? Focusing on the average ranked point forecast accuracy, we review the most commonly-used measures in both the academia and practice and perform a large-scale empirical study to understand the importance of the choice between measures. Our results suggest that there are only small discrepancies between the different error measures, especially within each measure category (percentage, relative, or scaled).

Kozak, S., Nagel, S., and Santosh, S. (2020). "Shrinking the cross-section." In: *Journal of Financial Economics* 135 (2), pp. 271–292.

We construct a robust stochastic discount factor (SDF) summarizing the joint explanatory power of a large number of cross-sectional stock return predictors. Our method achieves robust out-of-sample performance in this high-dimensional setting by imposing an economically motivated prior on SDF coefficients that shrinks contributions of low-variance principal components of the candidate characteristics-based factors. We find that characteristics-sparse SDFs formed from a few such factors.g., the four- or five-factor models in the recent literature adequately summarize the cross-section of expected stock returns. However, an SDF formed from a small number of principal components performs well.

Kruse, R., Leschinski, C., and Will, M. (2019). "Comparing Predictive Accuracy under Long Memory, With an Application to Volatility Forecasting." In: *Journal of Financial Econometrics* 17(2), pp. 180–228.

This article extends the popular Diebold-Mariano test for equal predictive accuracy to situations when the forecast error loss differential exhibits long memory. This situation can arise frequently since long memory can be transmitted from forecasts and the forecast objective to forecast error loss differentials. The nature of this transmission depends on the (un)biasedness of the forecasts and whether the involved series share common long memory. Further theoretical results show that the conventional Diebold-Mariano test is invalidated under these circumstances. Robust statistics based on a memory and autocorrelation consistent estimator and an extended fixed-bandwidth approach are considered. The subsequent extensive Monte Carlo study provides numerical results on various issues. As empirical applications, we consider recent extensions of the HAR model for the SandP500 realized volatility. While we find that forecasts improve significantly if jumps are considered, improvements achieved by the inclusion of an implied volatility index turn out to be insignificant.

Kuznetsov, V. and Mohri, M. (2016). "Time series prediction and online learning." In: *29th Annual Conference on Learning Theory*. Ed. by V. Feldman, A. Rakhlin, and O. Shamir. Vol. 49. Proceedings of Machine Learning Research. Columbia University, New York, New York, USA: PMLR, pp. 1190–1213.

We present a series of theoretical and algorithmic results combining the benefits of the statistical learning approach to time series prediction with that of on-line learning. We prove new generalization guarantees for hypotheses derived from regret minimization algorithms in the general scenario where the data is generated by a non-stationary non-mixing stochastic process. Our theory enables us to derive model selection techniques with favorable theoretical guarantees in the scenario of time series, thereby solving a problem that is notoriously difficult in that scenario. It also helps us devise new ensemble methods with favorable theoretical guarantees for the task of forecasting non-stationary time series.

Kynigakis, I. and Panopoulou, E. (2021). "Does Model Complexity add Value to Asset Allocation? Evidence from Machine Learning Forecasting Models." In: *Journal of Applied Econometrics*.

This study evaluates the benefits of integrating return forecasts from a variety of machine learning and forecast combination methods into an out-of-sample asset allocation framework. The economic evaluation of the forecasts shows that model complexity translates to improved results in the majority of cases considered, with shrinkage methods and shallow neural networks generating the highest individual performance. Overall, an investor would consistently realize superior out-of-sample gains by incorporating forecast combinations of machine learning models in the portfolio formation process.

Kyriakou, I., Mousavi, P., Nielsen, J. P., and Scholz, M. (2020). "Longer-Term Forecasting of Excess Stock Returns – The Five-Year Case." In: *Mathematics* 8(6), p. 927.

Long-term return expectations or predictions play an important role in planning purposes and guidance of long-term investors. Five-year stock returns are less volatile around their geometric mean than returns of higher frequency, such as one-year returns. One would, therefore, expect models using the latter to better reduce the noise and beat the simple historical mean than models based on the former. However, this paper shows that the general tendency is surprisingly the opposite: long-term forecasts over five years have a similar or even better predictive power when compared to the one-year case. We consider a long list of economic predictors and benchmarks relevant for the long-term investor. Our predictive approach consists of adopting and implementing a fully nonparametric smoother with the covariates and the smoothing parameters chosen by cross-validation. We consistently find that long-term forecasting performs well and recommend drawing more attention to it when designing investment strategies for long-term investors. Furthermore, our preferred predictive model did stand the test of Covid-19 providing a relatively optimistic outlook in March 2020 when uncertainty was all around us with lockdown and facing an unknown new pandemic.

Lara-Benitez, P., Carranza-Garcia, M., and Riquelme, J. C. (2021). "An Experimental Review on Deep Learning Architectures for Time Series Forecasting." In: *arXiv e-Print*.

In recent years, deep learning techniques have outperformed traditional models in many machine learning tasks. Deep neural networks have successfully been applied to address time series forecasting problems, which is a very important topic in data mining. They have proved to be an effective solution given their capacity to automatically learn the temporal dependencies present in time series. However, selecting the most convenient type of deep neural network and its parametrization is a complex task that requires considerable expertise. Therefore, there is a need for deeper studies on the suitability of all existing architectures for different forecasting tasks. In this work, we face two main challenges: a comprehensive review of the latest works using deep learning for time series forecasting; and an experimental study comparing the performance of the most popular architectures. The comparison involves a thorough analysis of seven types of deep learning models in terms of accuracy and efficiency. We evaluate the rankings and distribution of results obtained with the proposed models under many different architecture configurations and training hyperparameters. The datasets used comprise more than 50000 time series divided into 12 different forecasting problems. By training more than 38000 models on these data, we provide the most extensive deep learning study for time series forecasting. Among all studied models, the results show that long short-term memory (LSTM) and convolutional networks (CNN) are the best alternatives, with LSTMs obtaining the most accurate forecasts. CNNs achieve comparable performance with less variability of results under different parameter configurations, while also being more efficient.

Lara-Benítez, P., Gallego-Ledesma, L., Carranza-García, M., and Luna-Romera, J. M. (2021). "Evaluation of the Transformer Architecture for Univariate Time Series Forecasting." In: *Conference of the Spanish Association for Artificial IntelligenceCAEPIA 2021: Advances in Artificial Intelligence*. Springer International Publishing, pp. 106–115.

The attention-based Transformer architecture is earning increasing popularity for many machine learning tasks. In this study, we aim to explore the suitability of Transformers for time series forecasting, which is a crucial problem in different domains. We perform an extensive experimental study of the Transformer with different architecture and hyper-parameter configurations over 12 datasets with more than 50,000 time series. The forecasting accuracy and computational efficiency of Transformers are compared with state-of-the-art deep learning networks such as LSTM and CNN. The obtained results demonstrate that Transformers can outperform traditional recurrent or convolutional models due to their capacity to capture long-term dependencies, obtaining the most accurate forecasts in five out of twelve datasets. However, Transformers are generally more difficult to parametrize and show higher variability of results. In terms of efficiency, Transformer models proved to be less competitive in inference time and similar to the LSTM in training time.

Le Guen, V. and Thome, N. (2020). "Probabilistic Time Series Forecasting with Structured Shape and Temporal Diversity." In: *arXiv e-Print*.

Probabilistic forecasting consists in predicting a distribution of possible future outcomes. In this paper, we address this problem for non-stationary time series, which is very challenging yet crucially important. We introduce the STRIPE model for representing structured diversity based on shape and time features, ensuring both probable predictions while being sharp and accurate. STRIPE is agnostic to the forecasting model, and we equip it with a diversification mechanism relying on determinantal point processes (DPP). We introduce two DPP kernels for modeling diverse trajectories in terms of shape and time, which are both differentiable and proved to be positive semi-definite. To have an explicit control on the diversity structure, we also design an iterative sampling mechanism to disentangle shape and time representations in the latent space. Experiments carried out on synthetic datasets show that STRIPE significantly outperforms baseline methods for representing diversity, while maintaining accuracy of the forecasting model. We also highlight the relevance of the iterative sampling scheme and the importance to use different criteria for measuring quality and diversity. Finally, experiments on real datasets illustrate that STRIPE is able to outperform state-of-the-art probabilistic forecasting approaches in the best sample prediction.

Le Guen, V. and Thome, N. (2021). "Deep Time Series Forecasting with Shape and Temporal Criteria." In: *arXiv e-Print*.

This paper addresses the problem of multi-step time series forecasting for non-stationary signals that can present sudden changes. Current state-of-the-art deep learning forecasting methods, often trained with variants of the MSE, lack the ability to provide sharp predictions in deterministic and probabilistic contexts. To handle these challenges, we propose to incorporate shape and temporal criteria in the training objective of deep models. We define shape and temporal similarities and dissimilarities, based on a smooth relaxation of Dynamic Time Warping (DTW) and Temporal Distortion Index (TDI), that enable to build differentiable loss functions and positive semi-definite (PSD) kernels. With these tools, we introduce DILATE (DIstortion Loss including shApe and TimE), a new objective for deterministic forecasting, that explicitly incorporates two terms supporting precise shape and temporal change detection. For probabilistic forecasting, we introduce STRIPE++ (Shape and Time diverRsIty in Probabilistic forEcasting), a framework for providing a set of sharp and diverse forecasts, where the structured shape and time diversity is enforced with a determinantal point process (DPP) diversity loss. Extensive experiments and ablations studies on synthetic and real-world datasets confirm the benefits of leveraging shape and time features in time series forecasting.

Ledoit, O., Wolf, M., and Zhao, Z. (2019). "Efficient Sorting: A More Powerful Test for Cross-Sectional Anomalies." In: *Journal of Financial Econometrics* 17(4), pp. 645–686.

Many researchers seek factors that predict the cross-section of stock returns. The standard methodology sorts stocks according to their factor scores into quantiles and forms a corresponding long-short portfolio. Such a course of action ignores any information on the covariance matrix of stock returns. Historically, it has been difficult to estimate the covariance matrix for a large universe of stocks. We demonstrate that using the recent DCC-NL estimator of Engle, Ledoit, and Wolf (2017) substantially enhances the power of tests for cross-sectional anomalies: On average, t-statistics more than double.

Lee, T.-H. and Seregina, E. (2022). "Optimal Portfolio Using Factor Graphical Lasso." In: *arXiv e-Print*.

Graphical models are a powerful tool to estimate a high-dimensional inverse covariance (precision) matrix, which has been applied for a portfolio allocation problem. The assumption made by these models is a sparsity of the precision matrix. However, when stock returns are driven by common factors, such assumption does not hold. We address this limitation and develop a framework, Factor Graphical Lasso (FGL), which integrates graphical models with the factor structure in the context of portfolio allocation by decomposing a precision matrix into low-rank and sparse components. Our theoretical results and simulations show that FGL consistently estimates the portfolio weights and risk exposure and also that FGL is robust to heavy-tailed distributions which makes our method suitable for financial applications. FGL-based portfolios are shown to exhibit superior performance over several prominent competitors including equal-weighted and Index portfolios in the empirical application for the S&P500 constituents.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F., and Gneiting, T. (2015). "Forecaster's Dilemma: Extreme Events and Forecast Evaluation." In: *arXiv e-Print*.

In public discussions of the quality of forecasts, attention typically focuses on the predictive performance in cases of extreme events. However, the restriction of conventional forecast evaluation methods to subsets of extreme observations has unexpected and undesired effects, and is bound to discredit skillful forecasts when the signal-to-noise ratio in the data generating process is low. Conditioning on outcomes is incompatible with the theoretical assumptions of established forecast evaluation methods, thereby confronting forecasters with what we refer to as the forecaster's dilemma. For probabilistic forecasts, proper weighted scoring rules have been proposed as decision theoretically justifiable alternatives for forecast evaluation with an emphasis on extreme events. Using theoretical arguments, simulation experiments, and a real data study on probabilistic forecasts of U.S. inflation and gross domestic product growth, we illustrate and discuss the forecaster's dilemma along with potential remedies.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F., and Gneiting, T. (2017). "Forecaster's Dilemma: Extreme Events and Forecast Evaluation." In: *Statistical Science* 32(1), pp. 106–127.

In public discussions of the quality of forecasts, attention typically focuses on the predictive performance in cases of extreme events. However, the restriction of conventional forecast evaluation methods to subsets of extreme observations has unexpected and undesired effects, and is bound to discredit skillful forecasts when the signal-to-noise ratio in the data generating process is low. Conditioning on outcomes is incompatible with the theoretical assumptions of established forecast evaluation methods, thereby confronting forecasters with what we refer to as the forecaster's dilemma. For probabilistic forecasts, proper weighted scoring rules have been proposed as decision-theoretically justifiable alternatives for forecast evaluation with an emphasis on extreme events. Using theoretical arguments, simulation experiments and a real data study on probabilistic forecasts of U.S. inflation and gross domestic product (GDP) growth, we illustrate and discuss the forecaster's dilemma along with potential remedies.

Leroy, A., Latouche, P., Guedj, B., and Gey, S. (2020). "MAGMA: Inference and Prediction with Multi-Task Gaussian Processes." In: *arXiv e-Print*.

We investigate the problem of multiple time series forecasting, with the objective to improve multiple-step-ahead predictions. We propose a multi-task Gaussian process framework to simultaneously model batches of individuals with a common mean function and a specific covariance structure. This common mean is defined as a Gaussian process for which the hyper-posterior distribution is tractable. Therefore an EM algorithm can be derived for simultaneous hyper-parameters optimisation and hyper-posterior computation. Unlike previous approaches in the literature, we account for uncertainty and handle uncommon grids of observations while maintaining explicit formulations, by modelling the mean process in a non-parametric probabilistic framework. We also provide predictive formulas integrating this common mean process. This approach greatly improves the predictive performance far from observations, where information shared across individuals provides a relevant prior mean. Our overall algorithm is called Magma (standing for Multi tAsk Gaussian processes with common MeAn), and publicly available as a R package. The quality of the mean process estimation, predictive performances, and comparisons to alternatives are assessed in various simulated scenarios and on real datasets.

Lettau, M. and Pelger, M. (2020). "Factors That Fit the Time Series and Cross-Section of Stock Returns." In: *The Review of Financial Studies* 33(5), pp. 2274–2325.

We propose a new method for estimating latent asset pricing factors that fit the time series and cross-section of expected returns. Our estimator generalizes principal component analysis (PCA) by including a penalty on the pricing error in expected returns. Our approach finds weak factors with high Sharpe ratios that PCA cannot detect. We discover five factors with economic meaning that explain well the cross-section and time series of characteristic-sorted portfolio returns. The out-of-sample maximum Sharpe ratio of our factors is twice as large as with PCA with substantially smaller pricing errors. Our factors imply that a significant amount of characteristic information is redundant. Authors have furnished an Internet Appendix, which is available on the Oxford University Press Web site next to the link to the final published paper online. Code and data available at `https://mpelger.people.stanford.edu/data-and-code`.

Leung, E., Lohre, H., Mischlich, D., Shea, Y., and Stroh, M. (2020). "The Promises and Pitfalls of Machine Learning for Predicting Cross-Sectional Stock Returns." In: *SSRN e-Print*.

Recent research suggests that machine learning models dominate traditional linear models in predicting cross-sectional stock returns. Indeed, we confirm this finding when predicting one-month forward looking returns based on a set of common equity factors, including predictors such as short-term reversal. Despite this statistical advantage of machine learning model predictions, we demonstrate economic gains to be more limited and critically dependent on the ability to take risk and implement trades efficiently. Unlike traditional models, machine-learning models have struggled less over the last decade in discerning valuable predictions from cross-sectional equity characteristics.

Leung, E., Lohre, H., Mischlich, D., Shea, Y., and Stroh, M. (2021). "The Promises and Pitfalls of Machine Learning for Predicting Stock Returns." In: *The Journal of Financial Data Science* 3(2), pp. 21–50.

Recent research suggests that machine learning models dominate traditional linear models in predicting cross-sectional stock returns. The authors confirm this finding when predicting one-month-forward-looking returns based on a set of common stock characteristics, including predictors such as short-term reversal. Despite the statistical advantage of machine learning model predictions, the authors demonstrate that the economic gains tend to be more limited and critically dependent on the ability to take risk and implement trades efficiently. Unlike traditional models, machine learning models have been somewhat more effective over the past decade at discerning valuable predictions from cross-sectional equity characteristics.

Li, A. W. and Bastos, G. S. (2020). "Stock Market Forecasting Using Deep Learning and Technical Analysis: A Systematic Review." In: *IEEE Access* 8, pp. 185232–185242.
Stock market forecasting is one of the biggest challenges in the financial market since its time series has a complex, noisy, chaotic, dynamic, volatile, and non-parametric nature. However, due to computing development, an intelligent model can help investors and professional analysts reduce the risk of their investments. As Deep Learning models have been extensively studied in recent years, several studies have explored these techniques to predict stock prices using historical data and technical indicators. However, as the objective is to generate forecasts for the financial market, it is essential to validate the model through profitability metrics and model performance. Therefore, this systematic review focuses on Deep Learning models implemented for stock market forecasting using technical analysis. Discussions were made based on four main points of view: predictor techniques, trading strategies, profitability metrics, and risk management. This study showed that the LSTM technique is widely applied in this scenario (73.5%). This work significant contribution is to highlight some limitations found in the literature, such as only 35.3% of the studies analysed profitability, and only two articles implemented risk management. Therefore, despite the widely explored theme, there are still interesting open areas for research and development.

Li, L., Kang, Y., and Li, F. (2021). "Bayesian forecast combination using time-varying features." In: *arXiv e-Print*.
In this work, we propose a novel framework for density forecast combination by constructing time-varying weights based on time series features, which is called Feature-based Bayesian Forecasting Model Averaging (FEBAMA). Our framework estimates weights in the forecast combination via Bayesian log predictive scores, in which the optimal forecasting combination is determined by time series features from historical information. In particular, we use an automatic Bayesian variable selection method to add weight to the importance of different features. To this end, our approach has better interpretability compared to other black-box forecasting combination schemes. We apply our framework to stock market data and M3 competition data. Based on our structure, a simple maximum-a-posteriori scheme outperforms benchmark methods, and Bayesian variable selection can further enhance the accuracy for both point and density forecasts.

Lichtendahl, K. C. and Winkler, R. L. (2020). "Why do some combinations perform better than others?" In: *International Journal of Forecasting* 36(1), pp. 142–149.
The evidence from the literature on forecast combination shows that combinations generally perform well. We discuss here how the accuracy and diversity of the methods being combined and the robustness of the combination rule can influence performance, and illustrate this by showing that a simple, robust combination of a subset of the nine methods used in the M4 competition best combination performs almost as well as that forecast, and is easier to implement. We screened out methods with low accuracy or highly correlated errors and combined the remaining methods using a trimmed mean. We also investigated the accuracy risk (the risk of a bad forecast), proposing two new accuracy measures for this purpose. Our trimmed mean and the trimmed mean of all nine methods both had lower accuracy risk than either the best combination in the M4 competition or the simple mean of the nine methods.

Lim, B. and Zohren, S. (2021). "Time-series forecasting with deep learning: a survey." In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 379(2194), p. 20200209.
Numerous deep learning architectures have been developed to accommodate the diversity of time series datasets across different domains. In this article, we survey common encoder and decoder designs used in both one-step-ahead and multi-horizon time series forecasting - describing how temporal information is incorporated into predictions by each model. Next, we highlight recent developments in hybrid deep learning models, which combine well-studied statistical models with neural network components to improve pure methods in either category. Lastly, we outline some ways in which deep learning can also facilitate decision support with time series data.

Liu, M., Zeng, A., Xu, Z., Lai, Q., and Xu, Q. (2021a). "Time Series is a Special Sequence: Forecasting with Sample Convolution and Interaction." In: *arXiv e-Print*.
Time series is a special type of sequence data, a set of observations collected at even time intervals and ordered chronologically. Existing deep learning techniques use generic sequence models (e.g., recurrent neural network, Transformer model, or temporal convolutional network) for time series analysis, which ignore some of its unique properties. In particular, three components characterize time series: trend, seasonality, and irregular components, and the former two components enable us to perform forecasting with reasonable accuracy. Other types of sequence data do not have such characteristics. Motivated by the above, in this paper, we propose a novel neural network architecture that conducts sample convolution and interaction for temporal modeling and apply

it for the time series forecasting problem, namely **SCINet**. Compared to conventional dilated causal convolution architectures, the proposed downsample-convolve-interact architecture enables multi-resolution analysis besides expanding the receptive field of the convolution operation, which facilitates extracting temporal relation features with enhanced predictability. Experimental results show that SCINet achieves significant prediction accuracy improvement over existing solutions across various real-world time series forecasting datasets.

Liu, Z., Zhu, Z., Gao, J., and Xu, C. (2021b). "Forecast Methods for Time Series Data: A Survey." In: *IEEE Access* 9, pp. 91896–91912.

Research on forecasting methods of time series data has become one of the hot spots. More and more time series data are produced in various fields. It provides data for the research of time series analysis method, and promotes the development of time series research. Due to the generation of highly complex and large-scale time series data, the construction of forecasting models for time series data brings greater challenges. The main challenges of time series modeling are high complexity of time series data, low accuracy and poor generalization ability of prediction model. This paper attempts to cover the existing modeling methods for time series data and classify them. In addition, we make comparisons between different methods and list some potential directions for time series forecasting.

Loning, M., Bagnall, A., Ganesh, S., Kazakov, V., Lines, J., and Kiraly, F. J. (2019). "sktime: A Unified Interface for Machine Learning with Time Series." In: *arXiv e-Print*.

We present sktime – a new scikit-learn compatible Python library with a unified interface for machine learning with time series. Time series data gives rise to various distinct but closely related learning tasks, such as forecasting and time series classification, many of which can be solved by reducing them to related simpler tasks. We discuss the main rationale for creating a unified interface, including reduction, as well as the design of sktime's core API, supported by a clear overview of common time series tasks and reduction approaches.

Loning, M. and Kiraly, F. (2020). "Forecasting with sktime: Designing sktime's New Forecasting API and Applying It to Replicate and Extend the M4 Study." In: *arXiv e-Print*.

We present a new open-source framework for forecasting in Python. Our framework forms part of sktime, a machine learning toolbox with a unified interface for different time series learning tasks, like forecasting, but also time series classification and regression. We provide a dedicated forecasting interface, common statistical algorithms, and scikit-learn compatible tools for building composite machine learning models. We use sktime to both replicate key results from the M4 forecasting study and to extend it. sktime allows to easily build, tune and evaluate new models. We investigate the potential of common machine learning techniques for univariate forecasting, including reduction, boosting, ensembling, pipelining and tuning. We find that simple hybrid models can boost the performance of statistical models, and that pure machine learning models can achieve competitive forecasting performance on the hourly data sets, outperforming the statistical algorithms and coming close to the M4 winner model.

Makridakis, S., Hyndman, R. J., and Petropoulos, F. (2020). "Forecasting in social settings: The state of the art." In: *International Journal of Forecasting* 36 (1), pp. 15–28.

This paper provides a non-systematic review of the progress of forecasting in social settings. It is aimed at someone outside the field of forecasting who wants to understand and appreciate the results of the M4 Competition, and forms a survey paper regarding the state of the art of this discipline. It discusses the recorded improvements in forecast accuracy over time, the need to capture forecast uncertainty, and things that can go wrong with predictions. Subsequently, the review classifies the knowledge achieved over recent years into (i) what we know, (ii) what we are not sure about, and (iii) what we don knowIn the first two areas, we explore the difference between explanation and prediction, the existence of an optimal model, the performance of machine learning methods on time series forecasting tasks, the difficulties of predicting non-stable environments, the performance of judgment, and the value added by exogenous variables. The article concludes with the importance of (thin and) fat tails, the challenges and advances in causal inference, and the role of luck.

Makridakis, S. and Petropoulos, F. (2020). "The M4 competition: Conclusions." In: *International Journal of Forecasting* 36(1), pp. 224–227.

This M4 Competition special issue has brought together high caliber academics and top-level practitioners to comment on, discuss and criticize the M4 Competition, as well as the challenges facing the field and the best way forward. Their suggestions have been invaluable and have established a direct line of communication between the academic and business communities that we hope will continue to grow and strengthen. Similarly, we expect that the integration of the statistical and ML groups will be successful, providing a common effort to further advance the field. As a part of the wider data science field, forecasting is bound to play a critical role in future in identifying patterns in data and consequently forecasting as accurately as possible, while also providing realistic estimates of the uncertainty. We should add that this special issue contains detailed descriptions of the winning methods by their authors that can serve as starting points for using them, conducting additional academic research, and improving the practice of forecasting in business firms, while making sure that practitioners fully understand both its benefits and its limitations, as well as the fact that all future predictions are uncertain.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2019). "The M4 Competition: 100,000 time series and 61 forecasting methods." In: *International Journal of Forecasting* 36 (1), pp. 54–74.

The M4 Competition follows on from the three previous M competitions, the purpose of which was to learn from empirical evidence both how to improve the forecasting accuracy and how such learning could be used to advance the theory and practice of forecasting. The aim of M4 was to replicate and extend the three previous competitions by: (a) significantly increasing the number of series, (b) expanding the number of forecasting methods, and (c) including prediction intervals in the evaluation process as well as point forecasts. This paper covers all aspects of M4 in detail, including its organization and running, the presentation of its results, the top-performing methods overall and by categories, its major findings and their implications, and the computational requirements of the various methods. Finally, it summarizes its main conclusions and states the expectation that its series will become a testing ground for the evaluation of new methods and the improvement of the practice of forecasting, while also suggesting some ways forward for the field.

Mancuso, P., Piccialli, V., and Sudoso, A. M. (2021). "A machine learning approach for forecasting hierarchical time series." In: *Expert Systems with Applications* 182, p. 115102.

In this paper, we propose a machine learning approach for forecasting hierarchical time series. When dealing with hierarchical time series, apart from generating accurate forecasts, one needs to select a suitable method for producing reconciled forecasts. Forecast reconciliation is the process of adjusting forecasts to make them coherent across the hierarchy. In literature, coherence is often enforced by using a post-processing technique on the base forecasts produced by suitable time series forecasting methods. On the contrary, our idea is to use a deep neural network to directly produce accurate and reconciled forecasts. We exploit the ability of a deep neural network to extract information capturing the structure of the hierarchy. We impose the reconciliation at training time by minimizing a customized loss function. In many practical applications, besides time series data, hierarchical time series include explanatory variables that are beneficial for increasing the forecasting accuracy. Exploiting this further information, our approach links the relationship between time series features extracted at any level of the hierarchy and the explanatory variables into an end-to-end neural network providing accurate and reconciled point forecasts. The effectiveness of the approach is validated on three real-world datasets, where our method outperforms state-of-the-art competitors in hierarchical forecasting.

Martin, G. M., Loaiza-Maya, R., Frazier, D. T., Maneesoonthorn, W., and Hassan, A. R. (2020). "Optimal probabilistic forecasts: When do they work?" In: *arXiv e-Print*.

Proper scoring rules are used to assess the out-of-sample accuracy of probabilistic forecasts, with different scoring rules rewarding distinct aspects of forecast performance. Herein, we re-investigate the practice of using proper scoring rules to produce probabilistic forecasts that are 'optimal' according to a given score, and assess when their out-of-sample accuracy is superior to alternative forecasts, according to that score. Particular attention is paid to relative predictive performance under misspecification of the predictive model. Using numerical illustrations, we document several novel findings within this paradigm that highlight the important interplay between the true data generating process, the assumed predictive model and the scoring rule. Notably, we show that only when a predictive model is sufficiently compatible with the true process to allow a particular score criterion to reward what it is designed to reward, will this approach to forecasting reap benefits. Subject to this compatibility however, the superiority of the optimal forecast will be greater, the greater is the degree of misspecification. We explore these issues under a range of different scenarios, and using both artificially simulated and empirical data.

Martin, G. M., Loaiza-Maya, R., Maneesoonthorn, W., Frazier, D. T., and Ramírez-Hassan, A. (2022). "Optimal probabilistic forecasts: When do they work?" In: *International Journal of Forecasting*.

Proper scoring rules are used to assess the out-of-sample accuracy of probabilistic forecasts, with different scoring rules rewarding distinct aspects of forecast performance. Herein, we re-investigate the practice of using proper scoring rules to produce probabilistic forecasts that are 'optimal' according to a given score and assess when their out-of-sample accuracy is superior to alternative forecasts, according to that score. Particular attention is paid to relative predictive performance under misspecification of the predictive model. Using numerical illustrations, we document several novel findings within this paradigm that highlight the important interplay between the true data generating process, the assumed predictive model and the scoring rule. Notably, we show that only when a predictive model is sufficiently compatible with the true process to allow a particular score criterion to reward what it is designed to reward, will this approach to forecasting reap benefits. Subject to this compatibility, however, the superiority of the optimal forecast will be greater, the greater is the degree of misspecification. We explore these issues under a range of different scenarios and using both artificially simulated and empirical data.

Martinez, A. B., Castle, J. L., and Hendry, D. F. (2020). *Smooth Robust Multi-Horizon Forecasts*. Tech. rep. George Washington University.

We investigate whether smooth robust methods for forecasting can help mitigate pronounced and persistent failure across multiple forecast horizons. We demonstrate that naive predictors are interpretable as local estimators of the long-run relationship with the advantage of adapting quickly after a break, but at a cost of additional forecast error variance. Smoothing over naive estimates helps retain these advantages while reducing the costs, especially for longer forecast horizons. We derive the performance of these predictors after a location shift, and confirm the results using simulations. We apply smooth methods to forecasts of U.K. productivity and U.S. 10-year Treasury yields and show that they can dramatically reduce persistent forecast failure exhibited by forecasts from macroeconomic models and professional forecasters.

Masini, R. P., Medeiros, M. C., and Mendes, E. F. (2021). "Machine Learning Advances for Time Series Forecasting." In: *arXiv e-Print*.

In this paper we survey the most recent advances in supervised machine learning and high-dimensional models for time series forecasting. We consider both linear and nonlinear alternatives. Among the linear methods we pay special attention to penalized regressions and ensemble of models. The nonlinear methods considered in the paper include shallow and deep neural networks, in their feed-forward and recurrent versions, and tree-based methods, such as random forests and boosted trees. We also consider ensemble and hybrid models by combining ingredients from different alternatives. Tests for superior predictive ability are briefly reviewed. Finally, we discuss application of machine learning in economics and finance and provide an illustration with high-frequency financial data.

McCracken, M. W. (2020). "Tests of Conditional Predictive Ability: Existence, Size, and Power." In: *SSRN e-Print*.

We investigate a test of conditional predictive ability described in Giacomini and White (2006; Econometrica). Our main goal is simply to demonstrate existence of the nullhypothesis and, in doing so, clarify just how unlikely it is for this hypothesis to hold. We do so using a simple example of point forecasting under quadratic loss. We then provide simulation evidence on the size and power of the test. While the test can be accurately sized we find that power is typically low.

McDonald, S., Coleman, S., McGinnity, T. M., Li, Y., and Belatreche, A. (2014). "A comparison of forecasting approaches for capital markets." In: *IEEE Conference on Computational Intelligence for Financial Engineering and Economics (CIFEr)*. London, UK: IEEE, pp. 32–39.

In recent years, machine learning algorithms have become increasingly popular in financial forecasting. Their flexible, data-driven nature makes them ideal candidates for dealing with complex financial data. This paper investigates the effectiveness of a number of machine learning algorithms, and combinations of these algorithms, at generating one-step ahead forecasts of a number of financial time series. We find that hybrid models consisting of a linear statistical model and a nonlinear machine learning algorithm are effective at forecasting future values of the series, particularly in terms of the future direction of the series.

McMillan, D. G. (2021a). "Forecasting sector stock market returns." In: *Journal of Asset Management* 22(4), pp. 291–300.

We seek to forecast sector stock returns using established predictor variables. Existing empirical evidence focuses on market level data, and thus, sector data provide fertile ground for research. In addition to in-sample predictive regressions, we consider recursive and rolling forecasts and whether such forecasts can be used successfully in a sector rotation portfolio. The results for ten sectors and eleven predictor variables highlight that two variables, the default return and stock return variance, have significant predictive power across the stock market series. Forecast results are also supportive of these series (especially the default return), which can outperform benchmark and alternative forecast models across a range of metrics. A sector rotation strategy based on these forecasts produces positive abnormal returns and a Sharpe ratio higher than the baseline model. An examination of the sectors at each rotation reveals that a small number of dominate in the constructed portfolios.

McMillan, D. G. (2021b). "Forecasting U.S. stock returns." In: *The European Journal of Finance* 27(1-2), pp. 86–109.

We forecast quarterly US stock returns using a breadth of forecast variables, methods and metrics, including linear and non-linear regressions, rolling and recursive techniques, forecast combinations and statistical and economic evaluation. Thus, extending research in terms of the range of predictor series and the scope of analysis. Consistent with much of literature, a broad view over the full set of predictor variables indicates that such models are unable to beat the historical mean model. However, nuances reveal forecast success varies according to how the forecasts are evaluated and over time. Results reveal that the term structure of interest rates consistently provides the preferred forecast performance, especially when evaluated using the Sharpe ratio. The purchasing managers index also consistently provides a strong forecast performance. Further results reveal that forecast combinations over the full set of variables do not outperform the preferred single variable forecasts, while an interest rate forecast combination subset does perform well. The success of the term structure and the purchasing managers index highlights the importance of, respectively, investor and firm expectations of future economic performance in providing valuable stock return forecasts and is consistent with asset pricing models that indicate movements in returns are conditioned by such expectations.

Meligkotsidou, L., Panopoulou, E., Vrontos, I. D., and Vrontos, S. D. (2021). "Out-of-sample equity premium prediction: a complete subset quantile regression approach." In: *The European Journal of Finance* 27(1-2), pp. 110–135.

This paper extends the complete subset linear regression framework to a quantile regression setting. We employ complete subset combinations of quantile forecasts in order to construct robust and accurate equity premium predictions. We show that our approach delivers statistically and economically significant out-of-sample forecasts relative to both the historical average benchmark, the complete subset mean regression approach and the single-variable quantile forecast combination approach. Our recursive algorithm that selects, in real time, the best complete subset for each predictive regression quantile succeeds in identifying the best subset in a time- and quantile-varying manner.

Menezes, A. G. and Mastelini, S. M. (2021). "MegazordNet: combining statistical and machine learning standpoints for time series forecasting." In: *arXiv e-Print*.

Forecasting financial time series is considered to be a difficult task due to the chaotic feature of the series. Statistical approaches have shown solid results in some specific problems such as predicting market direction and single-price of stocks; however, with the recent advances in deep learning and big data techniques, new promising options have arises to tackle financial time series forecasting. Moreover, recent literature has shown that employing a combination of statistics and machine learning may improve accuracy in the forecasts in comparison to single solutions. Taking into consideration the mentioned aspects, in this work, we proposed the MegazordNet, a framework that explores statistical features within a financial series combined with a structured deep learning model for time series forecasting. We evaluated our approach predicting the closing price of stocks in the S&P 500 using different metrics, and we were able to beat single statistical and machine learning methods.

Messmer, M. and Audrino, F. (2020). "The Lasso and the Factor Zoo - Expected Returns in the Cross-Section." In: *SSRN e-Print*.

We document that cross-sectional return predictions based on OLS and Lasso type linear methods contain no predictive power for large cap stocks over the last decades. Small and micro cap stocks are highly predictable throughout the entire sample. Based on the 68 firm characteristics (FC) included in our analysis, the variable selection step suggests a highly multi-dimensional return process. Additionally, our Monte Carlo simulations indicate advantages of Lasso type predictions over OLS in panel specifications with a low signal-to-noise ratio. The results are robust to various assumptions.

Mikeliani, R. and Kavlashvili, N. (2020). "Evaluation and comparison of machine learning and classical econometric AR model on financial time series data." MA thesis. University of Tartu.

This paper examines the effects of time series data behaviour on the predictive performance of classical econometric univariate autoregressive and machine learning autoregressive models. The research aims to understand which forecasting approach would perform better in extreme scenarios. Even though some empirical studies demonstrate the superiority of machine learning methods relative to classical econometric methods, it is still arguable under what conditions one method can be constantly better than the other. And if there are any cases when econometric models are preferable than machine learning. Data is derived from simulation, ensuring the presence of different outlier and error distributions in small and relatively larger samples. The simulation results show that the machine learning approach outperforms econometric models in most of the cases. However, the existence of outliers worsens the performance of machine learning on small datasets. Even with the presence of outliers, as the sample size grows, the result improves so much for machine learning that it dominates the econometric model. The same models were used to forecast with rolling sample approaches on real financial data.

Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J., and Talagala, T. S. (2020). "FFORMA: Feature-based Forecast Model Averaging." In: *International Journal of Forecasting* 36 (1), pp. 86–92.

We propose an automated method for obtaining weighted forecast combinations using time series features. The proposed approach involves two phases. First, we use a collection of time series to train a meta-model to assign weights to various possible forecasting methods with the goal of minimizing the average forecasting loss obtained from a weighted forecast combination. The inputs to the meta-model are features extracted from each series. In the second phase, we forecast new series using a weighted forecast combination where the weights are obtained from our previously trained meta-model. Our method outperforms a simple forecast combination, and outperforms all of the most popular individual methods in the time series forecasting literature. The approach achieved second position in the M4 competition.

Montero-Manso, P. and Hyndman, R. J. (2020). "Principles and Algorithms for Forecasting Groups of Time Series: Locality and Globality." In: *arXiv e-Print*.

Forecasting groups of time series is of increasing practical importance, e.g. forecasting the demand for multiple products offered by a retailer or server loads within a data center. The local approach to this problem considers each time series separately and fits a function or model to each series. The global approach fits a single function to all series. For groups of similar time series, global methods outperform the more established local methods. However, recent results show good performance of global models even in heterogeneous datasets. This suggests a more general applicability of global methods, potentially leading to more accurate tools and new scenarios to study. Formalizing the setting of forecasting a set of time series with local and global methods, we provide the following contributions: 1) Global methods are not more restrictive than local methods, both can produce the same forecasts without any assumptions about similarity of the series. Global models can succeed in a wider range of problems than previously thought. 2) Basic generalization bounds for local and global algorithms. The complexity of local methods grows with the size of the set while it remains constant for global methods. In large datasets, a global algorithm can afford to be quite complex and still benefit from better generalization. These bounds serve to clarify and support recent experimental results in the field, and guide the design of new algorithms. For the class of autoregressive models, this implies that global models can have much larger memory than local methods. 3) In an extensive empirical study, purposely naive algorithms derived from these principles, such as global linear models or deep networks result in superior accuracy. In particular, global linear models can provide competitive accuracy with two orders of magnitude fewer parameters than local methods.

Montero-Manso, P. and Hyndman, R. J. (2021). "Principles and Algorithms for Forecasting Groups of Time Series: Locality and Globality." In: *International Journal of Forecasting* 37(4), pp. 1632–1653.

Global methods that fit a single forecasting method to all time series in a set have recently shown surprising accuracy, even when forecasting large groups of heterogeneous time series. We provide the following contributions that help understand the potential and applicability of global methods and how they relate to traditional local methods that fit a separate forecasting method to each series: Global and local methods can produce the same forecasts without any assumptions about similarity of the series in the set. The complexity of local methods grows with the size of the set while it remains constant for global methods. This result supports the recent evidence and provides principles for the design of new algorithms. In an extensive empirical study, we show that purposely naive algorithms derived from these principles show outstanding accuracy. In particular, global linear models provide competitive accuracy with far fewer parameters than the simplest of local methods.

Murray, M. H. and Blume, J. D. (2020). "False Discovery Rate Computation: Illustrations and Modifications." In: *arXiv e-Print*.

False discovery rates (FDR) are an essential component of statistical inference, representing the propensity for an observed result to be mistaken. FDR estimates should accompany observed results to help the user contextualize the relevance and potential impact of findings. This paper introduces a new user-friendly R package for computing FDRs and adjusting p-values for FDR control. These tools respect the critical difference between the adjusted p-value and the estimated FDR for a particular finding, which are sometimes numerically identical but are often confused in practice. Newly augmented methods for estimating the null proportion of findings - an important part of the FDR estimation procedure - are proposed and evaluated. The package is broad, encompassing a variety of methods for FDR estimation and FDR control, and includes plotting functions for easy display of results. Through extensive illustrations, we strongly encourage wider reporting of false discovery rates for observed findings.

Neri, F. (2021). "Domain Specific Concept Drift Detectors for Predicting Financial Time Series." In: *arXiv e-Print*.
Concept drift detectors allow learning systems to maintain good accuracy on non-stationary data streams. Financial time series are an instance of non-stationary data streams whose concept drifts (market phases) are so important to affect investment decisions worldwide. This paper studies how concept drift detectors behave when applied to financial time series. General results are: a) concept drift detectors usually improve the runtime over continuous learning, b) their computational cost is usually a fraction of the learning and prediction steps of even basic learners, c) it is important to study concept drift detectors in combination with the learning systems they will operate with, and d) concept drift detectors can be directly applied to the time series of raw financial data and not only to the model's accuracy one. Moreover, the study introduces three simple concept drift detectors, tailored to financial time series, and shows that two of them can be at least as effective as the most sophisticated ones from the state of the art when applied to financial time series.

Neto, A. E. D., Gonzalo, J., and Pitarakis, J.-Y. (2021). "Uncovering regimes in out of sample forecast errors." In: *Oxford Bulletin of Economics and Statistics* 83(3), pp. 713–741.
We introduce a set of test statistics for assessing the presence of regimes in out of sample forecast errors produced by recursively estimated linear multiple predictive regressions. These predictive regressions can accommodate multiple predictors that are highly persistent with potentially different degrees of persistence. Our method is also designed to be robust to the chosen starting window size so as to avert data mining concerns. Our tests are shown to be consistent and to lead to null distributions that are free of nuisance parameters and hence robust to the degree of persistence of the predictors.

Nevasalmi, L. (2020). "Forecasting multinomial stock returns using machine learning methods." In: *The Journal of Finance and Data Science* 6, pp. 86–106.
In this paper, the daily returns of the S&P 500 stock market index are predicted using a variety of different machine learning methods. We propose a new multinomial classification approach to forecasting stock returns. The multinomial approach can isolate the noisy fluctuation around zero return and allows us to focus on predicting the more informative large absolute returns. Our in-sample and out-of-sample forecasting results indicate significant return predictability from a statistical point of view. Moreover, all the machine learning methods considered outperform the benchmark buy-and-hold strategy in a real-life trading simulation. The gradient boosting machine is the top-performer in terms of both the statistical and economic evaluation criteria.

Nevasalmi, L. (2022). "Recession forecasting with high-dimensional data." In: *Journal of Forecasting*.
In this paper, a large amount of different financial and macroeconomic variables are used to predict the U.S. recession periods. We propose a new cost-sensitive extension to the gradient boosting model which can take into account the class imbalance problem of the binary response variable. The class imbalance, caused by the scarcity of recession periods in our application, is a problem that is emphasized with high-dimensional datasets. Our empirical results show that the introduced cost-sensitive extension outperforms the traditional gradient boosting model in both in-sample and out-of-sample forecasting. Among the large set of candidate predictors, different types of interest rate spreads turn out to be the most important predictors when forecasting U.S. recession periods.

Nietert, B. and Otto, T. (2020). "Empirical asset pricing: economic significance and economic model evaluation." In: *SSRN e-Print*.
Harvey (2017) and The American Statistical Association (2016) point out that business decisions should not be based only on whether the p-value of an empirical model passes a specific threshold and that statistical significance (p-value) cannot measure the size of an effect or the importance of a result. In other words, for economic problems economic significance is required and an economic model evaluation criterion is desirable.This paper derives a criterion for economic significance of valuation differences between empirical models and shows empirically that nearly all empirical models applied in business valuation are dis-similar, i.e., result in economically significant valuation differences. Motivated by the degree of dis-similarity between empirical models, an economic model evaluation criterion is developed. It judges the implicit economic assumptions revealed by computing the dual program of empirical models with the help of compliance with the economic principle and fit to institutional circumstances. Based on this economic model evaluation criterion our paper elaborates that within the group of cross-sectional price models quantile regression proves to be the best model because it is able to offer a good approximation to the economic principle and mimics best the institutional circumstances, in particular, if the regression is run without a constant. On the other hand, statistically more advanced models like generalized least squares regression deteriorates the implied economic content of models.

Noguer i Alonso, M., Batres-Estrada, G., and Moulin, A. (2020). "Deep Learning for Equity Time Series Prediction." In: *SSRN e-Print*.

We examine the performance of Deep Learning methods applied to equity financial time series. Predicting equity time series is a crucial topic in Finance. To form equity portfolios and do asset allocation, we need to predict returns, compute their risk, and optimize market impact. One of the modeling benefits of Deep Learning architectures is the ability to model non-linear highly dimensional problems. The lack of transparency and a rigorous mathematical theory could be considered less positive sides. The fact that most progress in Deep Learning has been made by trial and error is also cumbersome. Equity financial time series is a challenging domain with some stylized facts: weak stationarity, fat tails in return distributions, small data sets compared to other areas of Artificial Intelligence (AI), slow decay of autocorrelation in returns, and volatility clustering, to name the most important ones. We perform a comparative study between Long ShortTerm Memory Networks (LSTM), Recurrent Neural Networks (RNN), Deep Feed-Forward neural networks (DNN), and Gated Recurrent Unit Networks (GRU). We perform two types of studies. The first focused on a univariate test, and the second a multivariate test. Our tests show that the LSTM performs the best compared to other Deep Learning and classical machine learning models. In terms of performance metrics, the LSTM is better than the baseline model. We also show that the predictions are better than chance. There is enough evidence thatRNN and LSTM can deal with stationary time series and learn the data generating process. Nevertheless, predicting equity non-stationary time series, with market developments like the one caused by the COVID-19 pandemic in 2020, is challenging.

Noguer i Alonso, M. and Srivastava, S. (2021). "The Shape of Performance Curve in Financial Time Series." In: *SSRN e-Print*.

We examine in this paper a critical question in finance: the use of large nonlinear over-parametrized models or simpler models to forecast financial time series and the balance between underfitting and overfitting, the bias-variance trade-off, and the absolute performance in the test set. The traditional shape of the performance curve was U-shaped due to a bias-variance trade-off. Still, recently some recent research has pointed out that the performance curve may have a double descent shape in some specific domains.

We discuss some of the recent discoveries in the mathematical theory of machine learning that reduce the gap between theory and practice. We conduct experiments in the financial time series domain using deep neural networks and tree ensembles: random forests and XGBoost from under to over-parametrized.

The performance function doesn't show a U-shape or a double descent shape but a flat profile that means that larger models have the same performance in the test set than smaller models. However, the training error function shows a descent profile consistent with the idea that while training error can be very low when we increase the models' dimensionality, the test error is more stable in the equity financial time series domain. This is consistent with the finance practitioner's theory that backtesting (training data performance) frequently overestimates the test or real-life performance in financial time series prediction. The irreducible error limits the prediction performance.

Nonejad, N. (2021). "Bayesian model averaging and the conditional volatility process: an application to predicting aggregate equity returns by conditioning on economic variables." In: *Quantitative Finance* 21(8), pp. 1387–1411.

This study revisits the topic of predicting aggregate equity returns out-of-sample by conditioning on economic variables through Bayesian model averaging (BMA). Besides simultaneously addressing parameter instability and model uncertainty, I suggest a new model feature, namely, predictors in a given model can also impact the dependent variable through the conditional volatility process. The suggested econometric framework is straightforward to implement without requiring simulation. Likewise, the user can easily decide, which aspects of the predictive channel should to be switched on, off or altered. I apply the suggested framework to the well-known [Goyal, A. and Welch, I., A comprehensive look at the empirical performance of equity premium prediction. Rev. Financial Stud., 2008, 21, 1455-1508] dataset. An extensive out-of-sample prediction evaluation demonstrates that averaging over predictor combinations in a model that allows lagged predictors to impact aggregate equity returns exclusively through the conditional volatility process results in statistically significant more accurate density predictions relative to the benchmark, especially when predicting the left tail of the conditional distribution. One also observes economic gains in favor of certain BMAs. Here, the BMA that allows predictors to impact equity returns through the conditional mean as well as the conditional volatility process is the top performer.

Nybrant, A. (2021). "On Robust Forecast Combinations With Applications to Automated Forecasting." MA thesis. University of Uppsala.

Combining forecasts have been proven as one of the most successful methods to improve predictive performance. However, while there often is a focus on theoretically optimal methods, this is an ill-posed issue in practice where the problem of robustness is of more empirical relevance. This thesis focuses on the latter issue, where the risk associated with different combination methods is examined. The problem is addressed using Monte Carlo experiments and an application to automated forecasting with data from the M4 competition. Overall, our results indicate that the choice of combining methodology could constitute an important source of risk. While equal weighting of forecasts generally works well in the application, there are also cases where estimating weights improve upon this benchmark. In these cases, many robust and simple alternatives perform the best. While estimating weights can be beneficial, it is important to acknowledge the role of estimation uncertainty as it could outweigh the benefits of combining. For this reason, it could be advantageous to consider methods that effectively acknowledge this source of risk. By doing so, a forecaster can effectively utilize the benefits of combining forecasts while avoiding the risk associated with uncertainty in weights.

Nystrup, P., Lindstrom, E., Møller, J. K., and Madsen, H. (2021). "Dimensionality reduction in forecasting with temporal hierarchies." In: *International Journal of Forecasting* 37(3), pp. 1127–1146.

Combining forecasts from multiple temporal aggregation levels exploits information differences and mitigates model uncertainty, while reconciliation ensures a unified prediction that supports aligned decisions at different horizons. It can be challenging to estimate the full cross-covariance matrix for a temporal hierarchy, which can easily be of very large dimension, yet it is difficult to know a priori which part of the error structure is most important. To address these issues, we propose to use eigendecomposition for dimensionality reduction when reconciling forecasts to extract as much information as possible from the error structure given the data available. We evaluate the proposed estimator in a simulation study and demonstrate its usefulness through applications to short-term electricity load and financial volatility forecasting. We find that accuracy can be improved uniformly across all aggregation levels, as the estimator achieves state-of-the-art accuracy while being applicable to hierarchies of all sizes.

Odendahl, F., Rossi, B., and Sekhposyan, T. (2020). "Comparing Forecast Performance with State Dependence." In: *SSRN e-Print*.

We propose a novel forecast comparison methodology to evaluate models' relative forecasting performance when the latter is a state-dependent function of economic variables. In our benchmark case, the relative forecasting performance, measured by the forecast loss differential, is modeled via a threshold model. Importantly, we allow the threshold that triggers the switch from one state to the next to be unknown, leading to a non-standard test statistic due to the presence of a nuisance parameter. Existing tests either assume a constant out-of-sample forecast performance or use non-parametric techniques robust to time-variation; consequently, they may lack power against state-dependent predictability. Importantly, our approach is applicable to point forecasts as well as predictive densities. Monte Carlo results suggest that our proposed test statistics perform well in finite samples and have better power than existing tests in selecting the best forecasting model in the presence of state dependence. Our test statistics uncover "pockets of predictability" in U.S. equity premia forecasts; the pockets are a state-dependent function of stock market volatility. Models using economic predictors perform significantly worse than a simple mean forecast in periods of high volatility, but, in periods of low volatility, the use of economic predictors may lead to small forecast improvements.

Oh, D. H. and Patton, A. J. (2021). "Better the Devil You Know: Improved Forecasts from Imperfect Models." In: *Finance and Economics Discussion Series* 2021(070), pp. 1–45.

Many important economic decisions are based on a parametric forecasting model that is known to be good but imperfect. We propose methods to improve out-of-sample forecasts from a misspecified model by estimating its parameters using a form of local M estimation (thereby nesting local OLS and local MLE), drawing on information from a state variable that is correlated with the misspecification of the model. We theoretically consider the forecast environments in which our approach is likely to offer improvements over standard methods, and we find significant forecast improvements from applying the proposed method across distinct empirical analyses including volatility forecasting, risk management, and yield curve forecasting.

Okuno, S., Aihara, K., and Hirata, Y. (2019). "Combining multiple forecasts for multivariate time series via state-dependent weighting." In: *Chaos* 29(3), p. 033128.

We present a model-free forecast algorithm that dynamically combines multiple forecasts using multivariate time series data. The underlying principle is based on the fact that forecast performance depends on the position in the state space. This property is exploited to weight multiple forecasts via a local loss function. Specifically, additional weights are assigned to appropriate forecasts depending on their positions in a state space reconstructed via delay coordinates. The function evaluates the forecast error discounted by the distance in the space, whereas most existing methods discount the error in relation to time. In addition, forecasts are selected with the function for each time step based on the existing multiview embedding approach; by contrast, the original multiview embedding selects the reconstructions in advance before starting the forecast. The proposed prediction method has the smallest mean squared error among conventional ensemble methods for the Rossler and the Lorenz 96I models. The results of comparison of the proposed method with conventional machine-learning methods using a flood forecast example indicate that the proposed method yields superior accuracy. We also demonstrate that the proposed method might even correctly forecast the maximum water level of rivers without any prior knowledge.

Oreshkin, B. N., Carpov, D., Chapados, N., and Bengio, Y. (2019). "N-BEATS: Neural basis expansion analysis for interpretable time series forecasting." In: *arXiv e-Print*.
We focus on solving the univariate times series point forecasting problem using deep learning. We propose a deep neural architecture based on backward and forward residual links and a very deep stack of fully-connected layers. The architecture has a number of desirable properties, being interpretable, applicable without modification to a wide array of target domains, and fast to train. We test the proposed architecture on the well-known M4 competition dataset containing 100k time series from diverse domains. We demonstrate state-of-the-art performance for two configurations of N-BEATS, improving forecast accuracy by 11% over a statistical benchmark and by 3% over last year's winner of the M4 competition, a domain-adjusted hand-crafted hybrid between neural network and statistical time series models. The first configuration of our model does not employ any time-series-specific components and its performance on the M4 dataset strongly suggests that, contrarily to received wisdom, deep learning primitives such as residual blocks are by themselves sufficient to solve a wide range of forecasting problems. Finally, we demonstrate how the proposed architecture can be augmented to provide outputs that are interpretable without loss in accuracy.

Oreshkin, B. N., Carpov, D., Chapados, N., and Bengio, Y. (2020). "N-BEATS: Neural basis expansion analysis for interpretable time series forecasting." In: *ICLR 2020 Conference*.
We focus on solving the univariate times series point forecasting problem using deep learning. We propose a deep neural architecture based on backward and forward residual links and a very deep stack of fully-connected layers. The architecture has a number of desirable properties, being interpretable, applicable without modification to a wide array of target domains, and fast to train. We test the proposed architecture on the well-known M4 competition dataset containing 100k time series from diverse domains. We demonstrate state-of-the-art performance for two configurations of N-BEATS, improving forecast accuracy by 11% over a statistical benchmark and by 3% over last year's winner of the M4 competition, a domain-adjusted hand-crafted hybrid between neural network and statistical time series models. The first configuration of our model does not employ any time-series-specific components and its performance on the M4 dataset strongly suggests that, contrarily to received wisdom, deep learning primitives such as residual blocks are by themselves sufficient to solve a wide range of forecasting problems. Finally, we demonstrate how the proposed architecture can be augmented to provide outputs that are interpretable without loss in accuracy.

Papaioannou, P., Talmon, R., Serafino, D. di, and Siettos, C. (2021). "Time Series Forecasting Using Manifold Learning." In: *arXiv e-Print*.
We address a three-tier numerical framework based on manifold learning for the forecasting of high-dimensional time series. At the first step, we embed the time series into a reduced low-dimensional space using a nonlinear manifold learning algorithm such as Locally Linear Embedding and Diffusion Maps. At the second step, we construct reduced-order regression models on the manifold, in particular Multivariate Autoregressive (MVAR) and Gaussian Process Regression (GPR) models, to forecast the embedded dynamics. At the final step, we lift the embedded time series back to the original high-dimensional space using Radial Basis Functions interpolation and Geometric Harmonics. For our illustrations, we test the forecasting performance of the proposed numerical scheme with four sets of time series: three synthetic stochastic ones resembling EEG signals produced from linear and nonlinear stochastic models with different model orders, and one real-world data set containing daily time series of 10 key foreign exchange rates (FOREX) spanning the time period 19/09/2001-29/10/2020. The forecasting performance of the proposed numerical scheme is assessed using the combinations of manifold learning, modelling and lifting approaches. We also provide a comparison with the Principal Component Analysis algorithm as well as with the naive random walk model and the MVAR and GPR models trained and implemented directly in the high-dimensional space.

Paranhos, L. (2021). "Predicting Inflation with Neural Networks." In: *arXiv e-Print*.

This paper applies neural network models to forecast inflation. The use of a particular recurrent neural network, the long-short term memory model, or LSTM, that summarizes macroeconomic information into common components is a major contribution of the paper. Results from an exercise with US data indicate that the estimated neural nets usually present better forecasting performance than standard benchmarks, especially at long horizons. The LSTM in particular is found to outperform the traditional feed-forward network at long horizons, suggesting an advantage of the recurrent model in capturing the long-term trend of inflation. This finding can be rationalized by the so called long memory of the LSTM that incorporates relatively old information in the forecast as long as accuracy is improved, while economizing in the number of estimated parameters. Interestingly, the neural nets containing macroeconomic information capture well the features of inflation during and after the Great Recession, possibly indicating a role for nonlinearities and macro information in this episode. The estimated common components used in the forecast seem able to capture the business cycle dynamics, as well as information on prices.

Patton, A. J. (2020). "Comparing Possibly Misspecified Forecasts." In: *Journal of Business & Economic Statistics* 38(4), pp. 796–809.

Recent work has emphasized the importance of evaluating estimates of a statistical functional (such as a conditional mean, quantile, or distribution) using a loss function that is consistent for the functional of interest, of which there is an infinite number. If forecasters all use correctly specified models free from estimation error, and if the information sets of competing forecasters are nested, then the ranking induced by a single consistent loss function is sufficient for the ranking by any consistent loss function. This article shows, via analytical results and realistic simulation-based analyses, that the presence of misspecified models, parameter estimation error, or nonnested information sets, leads generally to sensitivity to the choice of (consistent) loss function. Thus, rather than merely specifying the target functional, which narrows the set of relevant loss functions only to the class of loss functions consistent for that functional, forecast consumers or survey designers should specify the single specific loss function that will be used to evaluate forecasts. An application to survey forecasts of U.S. inflation illustrates the results.

Perron, P. and Yamamoto, Y. (2021). "Testing for Changes in Forecasting Performance." In: *Journal of Business & Economic Statistics* 39(1), pp. 148–165.

We consider the issue of forecast failure (or breakdown) and propose methods to assess retrospectively whether a given forecasting model provides forecasts which show evidence of changes with respect to some loss function. We adapt the classical structural change tests to the forecast failure context. First, we recommend that all tests should be carried with a fixed scheme to have best power. This ensures a maximum difference between the fitted in and out-of-sample means of the losses and avoids contamination issues under the rolling and recursive schemes. With a fixed scheme, Giacomini and Rossi's (GR) test is simply a Wald test for a one-time change in the mean of the total (the in-sample plus out-of-sample) losses at a known break date, say m, the value that separates the in and out-of-sample periods. To alleviate this problem, we consider a variety of tests: maximizing the GR test over values of m within a prespecified range; a Double sup-Wald (DSW) test which for each m performs a sup-Wald test for a change in the mean of the out-of-sample losses and takes the maximum of such tests over some range; we also propose to work directly with the total loss series to define the Total Loss sup-Wald and Total Loss UDmax (TLUD) tests. Using theoretical analyses and simulations, we show that with forecasting models potentially involving lagged dependent variables, the only tests having a monotonic power function for all data-generating processes considered are the DSW and TLUD tests, constructed with a fixed forecasting window scheme. Some explanations are provided and empirical applications illustrate the relevance of our findings in practice. Supplementary materials for this article are available online.

Pesaran, M. H. and Smith, R. (2019). "The Role of Factor Strength and Pricing Errors for Estimation and Inference in Asset Pricing Models." In: *SSRN e-Print*.

In this paper we are concerned with the role of factor strength and pricing errors in asset pricing models, and their implications for identification and estimation of risk premia. We establish an explicit relationship between the pricing errors and the presence of weak factors that are correlated with stochastic discount factor. We introduce a measure of factor strength, and distinguish between observed factors and unobserved factors. We show that unobserved factors matter for pricing if they are correlated with the discount factor, and relate the strength of the weak factors to the strength (pervasiveness) of non-zero pricing errors. We then show, that even when the factor loadings are known, the risk premia of a factor can be consistently estimated only if it is strong and if the pricing errors are weak. Similar results hold when factor loadings are estimated, irrespective of whether individual returns or portfolio returns are used. We derive distributional results for two pass estimators of risk premia, allowing for non-zero pricing errors. We show that for inference on risk premia the pricing errors must be sufficiently weak. We consider both when n (the number of securities) is large and T (the number of time periods) is short, and the case of large n and T. Large n is required for consistent estimation of risk premia, whereas the choice of short T is intended to reduce the possibility of time variations in the factor loadings. We provide monthly rolling estimates of the factor strengths for the three Fama-French factors over the period 1989-2018.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Oliveira, F. L. C., De Baets, S., Dokumentov, A., Ellison, J., Fiszeder, P., Franses, P. H., Frazier, D. T., Gilliland, M., Gonul, M. S., Goodwin, P., Grossi, L., Grushka-Cockayne, Y., Guidolin, M., Guidolin, M., Gunter, U., Guo, X., Guseo, R., Harvey, N., Hendry, D. F., Hollyman, R., Januschowski, T., Jeon, J., Jose, V. R. R., Kang, Y., Koehler, A. B., Kolassa, S., Kourentzes, N., Leva, S., Li, F., Litsiou, K., Makridakis, S., Martin, G. M., Martinez, A. B., Meeran, S., Modis, T., Nikolopoulos, K., Onkal, D., Paccagnini, A., Panagiotelis, A., Panapakidis, I., Pavia, J. M., Pedio, M., Pedregal, D. J., Pinson, P., Ramos, P., Rapach, D. E., Reade, J. J., Rostami-Tabar, B., Rubaszek, M., Sermpinis, G., Shang, H. L., Spiliotis, E., Syntetos, A. A., Talagala, P. D., Talagala, T. S., Tashman, L., Thomakos, D., Thorarinsdottir, T., Todini, E., Arenas, J. R. T., Wang, X., Winkler, R. L., Yusupova, A., and Ziel, F. (2022). "Forecasting: theory and practice." In: *International Journal of Forecasting*.

Petropoulos, F. and Grushka-Cockayne, Y. (2021). "Fast and Frugal Time Series Forecasting." In: *SSRN e-Print*.
Over the years, families of forecasting models, such as the exponential smoothing family and Autoregressive Integrated Moving Average, have expanded to contain multiple possible forms and forecasting profiles. In this paper, we question the need to consider such large families of models. We argue that parsimoniously identifying suitable subsets of models will not decrease the forecasting accuracy nor will it reduce the ability to estimate the forecast uncertainty. We propose a framework that balances forecasting performance versus computational cost, resulting in a set of reduced families of models and empirically demonstrate this trade-offs. We translate computational benefits to monetary cost savings and discuss the implications of our results in the context of large retailers.

Petropoulos, F. and Spiliotis, E. (2021). "The Wisdom of the Data: Getting the Most Out of Univariate Time Series Forecasting." In: *Forecasting* 3(3), pp. 478–497.
Forecasting is a challenging task that typically requires making assumptions about the observed data but also the future conditions. Inevitably, any forecasting process will result in some degree of inaccuracy. The forecasting performance will further deteriorate as the uncertainty increases. In this article, we focus on univariate time series forecasting and we review five approaches that one can use to enhance the performance of standard extrapolation methods. Much has been written about the ”wisdom of the crowds” and how collective opinions will outperform individual ones. We present the concept of the ”wisdom of the data” and how data manipulation can result in information extraction which, in turn, translates to improved forecast accuracy by aggregating (combining) forecasts computed on different perspectives of the same data. We describe and discuss approaches that are based on the manipulation of local curvatures (theta method), temporal aggregation, bootstrapping, sub-seasonal and incomplete time series. We compare these approaches with regards to how they extract information from the data, their computational cost, and their performance.

Petropoulos, F., Spiliotis, E., and Panagiotelis, A. (2021). "Model combinations through revised base-rates." In: *arXiv e-Print*.
Standard selection criteria for forecasting models focus on information that is calculated for each series independently, disregarding the general tendencies and performances of the candidate models. In this paper, we propose a new way to statistical model selection and model combination that incorporates the base-rates of the candidate forecasting models, which are then revised so that the per-series information is taken into account. We examine two schemes that are based on the precision and sensitivity information from the contingency table of the base rates. We apply our approach on pools of exponential smoothing models and a large number of real time series and we show that our schemes work better than standard statistical benchmarks. We discuss the connection of our approach to other cross-learning approaches and offer insights regarding implications for theory and practice.

Petropoulos, F. and Svetunkov, I. (2020). "A simple combination of univariate models." In: *International Journal of Forecasting* 36(1), pp. 110–115.

Abstract This paper describes the approach that we implemented for producing the point forecasts and prediction intervals for our M4-competition submission. The proposed simple combination of univariate models (SCUM) is a median combination of the point forecasts and prediction intervals of four models, namely exponential smoothing, complex exponential smoothing, automatic autoregressive integrated moving average and dynamic optimised theta. Our submission performed very well in the M4-competition, being ranked 6th for the point forecasts (with a small difference compared to the 2nd submission) and prediction intervals and 2 nd and 3rd for the point forecasts of the weekly and quarterly data respectively.

Petropoulos, F., Wang, X., and Disney, S. M. (2019). "The inventory performance of forecasting methods: Evidence from the M3 competition data." In: *International Journal of Forecasting* 35(1), pp. 251–265.

Forecasting competitions have been a major driver not only of improvements in forecasting methods performances, but also of the development of new forecasting approaches. However, despite the tremendous value and impact of these competitions, they do suffer from the limitation that performances are measured only in terms of the forecast accuracy and bias, ignoring utility metrics. Using the monthly industry series of the M3 competition, we empirically explore the inventory performances of various widely used forecasting techniques, including exponential smoothing, ARIMA models, the Theta method, and approaches based on multiple temporal aggregation. We employ a rolling simulation approach and analyse the results for the order-up-to policy under various lead times. We find that the methods that are based on combinations result in superior inventory performances, while the Naive, Holt, and Holt-Winters methods perform poorly.

Pinho, D. M. (2020). "Forecast comparison of volatility models and their combinations (FTSE100): a tied race." MA thesis. Universidade do Minho.

I first compare 74 models that include the main naive, ARMA, GARCH, and HAR models from the volatility forecasting literature to assess their out-of-sample performance for day- ahead forecasts. For the FTSE100 index in the period of 2005-2010, all HAR and GARCH models and some ARMA and exponential smoothing models perform similarly to each other. I then test 176 model combinations (meaning that 250 models are compared in total) in the period of 2007-2010, and observe that the average performance has less variance and is always slightly improved. This tendency is not observed with complex weighting schemes, which are based on regularized regression (i.e., Lasso, Ridge and Elastic Net); and this tendency is marginally larger when excluding underperforming models and equal weighting the forecasts of the remaining models (i.e., trimming). But, overall, as observed in the relevant literature, all reasonably adequate models tend to have identical performance, so past research seems to have overstated the improvements generated by new models. An additional problem is that, according to the literature, even large performance gains with the loss functions used seem to rarely translate into improvements in economic applica- tions, such as risk management and portfolio optimization. Because of this, I argue that subsequent research must use metrics directly related to these applications.

Pinto, J. M. and Castle, J. (2021). *A machine learning dynamic switching approach to forecasting when there are structural breaks*. Tech. rep. University of Oxford.

Forecasting economic indicators is an important task for analysts. However, many indicators suffer from structural breaks leading to forecast failure. Methods that are robust following a structural break have been proposed in the literature but they come at a cost: an increase in forecast error variance. We propose a method to select between a set of robust and non-robust forecasting models. Our method uses time-series clustering to identify possible structural breaks in a time series, and then switches between forecasting models depending on the series dynamics. We perform a rigorous empirical evaluation with 400 simulated series with an artificial structural break and with real data economic series: Industrial Production and Consumer Prices for all Western European countries available from the OECD database. Our results show that the proposed method statistically outperforms benchmarks in forecast accuracy for most case scenarios, particularly at short horizons.

Pinto, J. M. and Marçal, E. F. (2019). "Cross-Validation Based Forecasting Method: A Machine Learning Approach." In: *SSRN e-Print*.

Our paper aims to evaluate two novel methods on selecting the best forecasting model or its combination based on a Machine Learning approach. The methods are based on the selection of the "best" model, or combination of models, by cross-validation technique, from a set of possible models. The first one is based on the seminal paper of Granger-Bates (1969) but weights are estimated by a process of cross-validation applied on the training set. The second one selects the model with the best forecasting performance in the process described above, which we called CvML (Cross-Validation Machine Learning Method). The following models are used: exponential smoothing, SARIMA, artificial neural networks and Threshold autoregression (TAR). Model specification is chosen by R packages: forecast and TSA. Both methods – CvML and MGB – are applied to these models to

generate forecasts from one up to twelve periods ahead. Frequency of data is monthly. We run the forecasts exercise to the following to monthly series of Industrial Product Indices for seven countries: Canada, Brazil, Belgium, Germany, Portugal, UK and USA. The data was collected at OECD data, with 504 observations. We choose Average Forecast Combination, Granger Bates Method, MCS model, Naive and Seasonal Naive Model as benchmarks. Our results suggest that MGB did not performed well. However, CvML had a lower mean absolute error for most of countries and forecast horizons, particularly at longer horizons, surpassing all the proposed benchmarks. Similar results hold for absolute mean forecast error.

Pinto, J. M. and Marçal, E. F. (2020). "Inflation Rate Forecasting: Extreme Learning Machine as a Model Combination Method." In: *Contributions to Statistics*. Springer International Publishing, pp. 365–385.

Inflation rate forecasting is one most discussed topics on time-series analysis due to its importance on macroeconomic policy. The majority of these papers' findings point out that forecasting combination methods usually outperform individual models. In this sense, we evaluate a novel method to combine forecasts based on Extreme Learning Machine Method [15], which is becoming very popular but, to the best of our knowledge, has not been used to this purpose. We test Inflation Rate forecasting for a set of American countries, for one, two, three, ten, eleven and twelve steps ahead. The models to be combined are automatically estimated by R forecast package, as SARIMA, Exponential Smoothing, ARFIMA, Spline Regression, and Artificial Neural Networks. Another goal of our paper is to test our model against classical combination methods such Granger Bates, Linear Regression, and Average Mean of models as benchmarks, but also test it against basic forms of new models in the literature, like [8, 10, 26]. Therefore, our paper also contributes to the discussion of forecast combination by comparing versions of some methods that have not been tested against each other. Our results indicate that none of these methods have an indisputable superiority against the others, however, the Extreme Learning Method proved to be the most efficient of all, with the smaller Mean Absolute Error and Mean Squared Error for its predictions.

Pitarakis, J.-Y. (2020). "A Novel Approach to Predictive Accuracy Testing in Nested Environments." In: *arXiv e-Print*.

We introduce a new approach for comparing the predictive accuracy of two nested models that bypasses the difficulties caused by the degeneracy of the asymptotic variance of forecast error loss differentials used in the construction of commonly used predictive comparison statistics. Our approach continues to rely on the out of sample MSE loss differentials between the two competing models, leads to nuisance parameter free Gaussian asymptotics and is shown to remain valid under flexible assumptions that can accommodate heteroskedasticity and the presence of mixed predictors (e.g. stationary and local to unit root). A local power analysis also establishes its ability to detect departures from the null in both stationary and persistent settings. Simulations calibrated to common economic and financial applications indicate that our methods have strong power with good size control across commonly encountered sample sizes.

Post, T., Karabati, S., and Arvanitis, S. (2019). "Robust optimization of forecast combinations." In: *International Journal of Forecasting* 35(3), pp. 910–926.

We develop a methodology for constructing robust combinations of time series forecast models which improve upon a given benchmark specification for all symmetric and convex loss functions. Under standard regularity conditions, the optimal forecast combination asymptotically almost surely dominates the benchmark, and also optimizes the chosen goal function. The optimum in a given sample can be found by solving a convex optimization problem. An application to the forecasting of changes in the S&P 500 volatility index shows that robust optimized combinations improve significantly upon the out-of-sample forecasting accuracy of both simple averaging and unrestricted optimization.

Prasad, V. V., Gumparthi, S., Venkataramana, L. Y., Srinethe, S., Sree, R. M. S., and Nishanthi, K. (2021). "Prediction of Stock Prices Using Statistical and Machine Learning Models: A Comparative Analysis." In: *The Computer Journal*.

With the advent of machine learning, numerous approaches have been proposed to forecast stock prices. Various models have been developed to date such as Recurrent Neural Networks, Long Short-Term Memory, Convolutional Neural Network sliding window, etc., but were not accurate enough. Here, the aim is to predict the price of a stock and compare the results obtained using three major algorithms namely Kalman filters, XGBoost and ARIMA. Kalman filters are recursive and use a feedback mechanism to perform error correction. This correction makes them best suited for making accurate predictions as they can factor in the market volatility, whereas XGBoost is a promising technique for datasets that are nonlinear and can gather knowledge by detecting patterns and relationships in the data. XGBoost is also capable of capturing the time dependency of features efficiently. ARIMA refers to an Auto Regressive Integrated Moving Average model that has become very popular in recent

times. It is mostly used on time series data and works by eliminating its stationarity. Finally, a hybrid model combining Kalman filters and XGBoostis discussed and a comparison of the results of each of the four models, are made to provide a better clarity for making investments by forecasting the price of a stock.

Prayogo, N., Cevik, M., and Bodur, M. (2020). "Time Series Sampling for Probabilistic Forecasting." In: *Proceedings of the 30th Annual International Conference on Computer Science and Software Engineering*. USA: IBM Corp., pp. 153–162.

Deep learning-based models for multiple time series probabilistic forecasting have gained significant attention in the recent literature. Given the abundance of data, many successful global and hybrid models that can learn complex patterns from multiple related time series have been developed. The main focus being on novel architecture development, little attention has been given to the investigation of input data for those models, making the impression that using more data is always better. In this paper, we strive to answer the following question: Is using more time series always better? Specifically, we investigate the usefulness of time series sampling in achieving better performance within lower run times on time series probabilistic forecasting. We evaluate the performance of two state-of-the-art models, namely DeepAR and DeepState, when using different numbers of time series that are selected based on a variety of distance-based similarity criteria for forecasting a single target time series. Through empirical evaluation on various common real-life datasets from the literature, we show that strategically selecting time series to train could help state-of-the-art models achieve improved forecasting accuracy while requiring a significantly less model training time.

Qian, W., Rolling, C. A., Cheng, G., and Yang, Y. (2019). "On the forecast combination puzzle." In: *Econometrics* 7(3), p. 39.

It is often reported in the forecast combination literature that a simple average of candidate forecasts is more robust than sophisticated combining methods. This phenomenon is usually referred to as the combination puzzle. Motivated by this puzzle, we explore its possible explanations, including high variance in estimating the target optimal weights (estimation error), invalid weighting formulas, and model/candidate screening before combination. We show that the existing understanding of the puzzle should be complemented by the distinction of different forecast combination scenarios known as combining for adaptation and combining for improvement. Applying combining methods without considering the underlying scenario can itself cause the puzzle. Based on our new understandings, both simulations and real data evaluations are conducted to illustrate the causes of the puzzle. We further propose a multi-level AFTER strategy that can integrate the strengths of different combining methods and adapt intelligently to the underlying scenario. In particular, by treating the simple average as a candidate forecast, the proposed strategy is shown to reduce the heavy cost of estimation error and, to a large extent, mitigate the puzzle.

Qian, W., Rolling, C. A., Cheng, G., and Yang, Y. (2022). "Combining forecasts for universally optimal performance." In: *International Journal of Forecasting*.

There are two potential directions of forecast combination: combining for adaptation and combining for improvement. The former direction targets the performance of the best forecaster, while the latter attempts to combine forecasts to improve on the best forecaster. It is often useful to infer which goal is more appropriate so that a suitable combination method may be used. This paper proposes an AI-AFTER approach that can not only determine the appropriate goal of forecast combination but also intelligently combine the forecasts to automatically achieve the proper goal. As a result of this approach, the combined forecasts from AI-AFTER perform well universally in both adaptation and improvement scenarios. The proposed forecasting approach is implemented in our R package AIafter, which is available at `https://github.com/weiqian1/AIafter`.

Qu, R., Timmermann, A., and Zhu, Y. (2021). "Comparing forecasting performance in cross-sections." In: *Journal of Econometrics*.

This paper develops new methods for pairwise comparisons of predictive accuracy with cross-sectional data. Using a common factor setup, we establish conditions on cross-sectional dependencies in forecast errors which allow us to test the null of equal predictive accuracy on a single cross-section of forecasts. We consider both unconditional tests of equal predictive accuracy as well as tests that condition on the realization of common factors and show how to decompose forecast errors into exposures to common factors and idiosyncratic components. An empirical application compares the predictive accuracy of financial analysts' short-term earnings forecasts across six brokerage firms.

Quaedvlieg, R. (2021). "Multi-Horizon Forecast Comparison." In: *Journal of Business & Economic Statistics*.

We introduce tests for multi-horizon superior predictive ability (SPA). Rather than comparing forecasts of different models at multiple horizons individually, we propose to jointly consider all horizons of a forecast path.

We define the concepts of uniform and average SPA. The former entails superior performance at each individual horizon, while the latter allows inferior performance at some horizons to be compensated by others. The article illustrates how the tests lead to more coherent conclusions, and how they are better able to differentiate between models than the single-horizon tests. We provide an extension of the previously introduced model confidence set to allow for multi-horizon comparison of more than two models. Simulations demonstrate appropriate size and high power. An illustration of the tests on a large set of macroeconomic variables demonstrates the empirical benefits of multi-horizon comparison.

Radchenko, P., Vasnev, A. L., and Wang, W. (2022). "Too similar to combine? On negative weights in forecast combination." In: *International Journal of Forecasting*.

This paper provides the first thorough investigation of the negative weights that can emerge when combining forecasts. The usual practice in the literature is to consider only convex combinations and ignore or trim negative weights, i.e., set them to zero. This default strategy has its merits, but it is not optimal. We study the problem from various angles, and the main conclusion is that negative weights emerge when highly correlated forecasts with similar variances are combined. In this situation, the estimated weights have large variances, and trimming reduces the variance of the weights and improves the combined forecast. The threshold of zero is arbitrary and can be improved. We propose an optimal trimming threshold, i.e., an additional tuning parameter to improve forecasting performance. The effects of optimal trimming are demonstrated in simulations. In the empirical example using the European Central Bank Survey of Professional Forecasters, we find that the new strategy performs exceptionally well and can deliver improvements of more than 10% for inflation, up to 20% for GDP growth, and more than 20% for unemployment forecasts relative to the equal-weight benchmark.

Rahimikia, E. and Poon, S.-H. (2021). "Machine Learning for Realised Volatility Forecasting." In: *SSRN e-Print*.

This paper compares machine learning and HAR models for forecasting realised volatility of 23 NASDAQ stocks using 146 variables extracted from limit order book (LOBSTER) and stock-specific news (Dow Jones Newswires) from 27 July 2007 to 18 November 2016. We find simpler ML to outperform HARs on normal volatility days. With SHAP, an Explainable AI technique, we find simple mid prices at all limit order book levels and mean bid/ask prices drive RV forecasts for many stocks. An ML model with a larger number of units and complex idiosyncratic LOB variables are needed for forecasting volatility jumps.

Rajapaksha, D., Bergmeir, C., and Hyndman, R. J. (2021). "LoMEF: A Framework to Produce Local Explanations for Global Model Time Series Forecasts." In: *arXiv e-Print*.

Global Forecasting Models (GFM) that are trained across a set of multiple time series have shown superior results in many forecasting competitions and real-world applications compared with univariate forecasting approaches. One aspect of the popularity of statistical forecasting models such as ETS and ARIMA is their relative simplicity and interpretability (in terms of relevant lags, trend, seasonality, and others), while GFMs typically lack interpretability, especially towards particular time series. This reduces the trust and confidence of the stakeholders when making decisions based on the forecasts without being able to understand the predictions. To mitigate this problem, in this work, we propose a novel local model-agnostic interpretability approach to explain the forecasts from GFMs. We train simpler univariate surrogate models that are considered interpretable (e.g., ETS) on the predictions of the GFM on samples within a neighbourhood that we obtain through bootstrapping or straightforwardly as the one-step-ahead global black-box model forecasts of the time series which needs to be explained. After, we evaluate the explanations for the forecasts of the global models in both qualitative and quantitative aspects such as accuracy, fidelity, stability and comprehensibility, and are able to show the benefits of our approach.

Rapach, D. and Zhou, G. (2022). "Asset Pricing: Time-Series Predictability." In: *SSRN e-Print*.

Asset returns change with fundamentals and other factors such as technical information and sentiment over time. This review covers some of the major ideas, data, and methods used to model time-varying expected returns. The focus is on the out-of-sample predictability of the aggregate stock market return via extensions of the conventional predictive regression approach.

The extensions are designed to improve out-of-sample performance in realistic environments characterized by large information sets and noisy data. Large information sets are relevant because a plethora of plausible stock return predictors exists. The information sets include variables typically associated with a rational time-varying market risk premium, as well as variables more likely to reflect market inefficiencies resulting from behavioral influences and information frictions. Noisy data stem from the intrinsically large unpredictable component in stock returns. When forecasting with large information sets and noisy data, it is vital to employ methods that incorporate the relevant information in the large set of predictors in a manner that guards against overfitting

the data.
Methods that improve out-of-sample market return prediction include forecast combination, principal component regression, partial least squares, the LASSO and elastic net from machine learning, and a newly developed C-ENet approach that relies on the elastic net to refine the simple combination forecast. Employing these methods, a number of studies provide statistically and economically significant evidence that the aggregate market return is predictable on an out-of-sample basis. Out-of-sample market return predictability based on a rich set of predictors thus appears to be a well-established empirical result in asset pricing.

Rapach, D. E., Strauss, J. K., Tu, J., and Zhou, G. (2019). "Industry return predictability: A machine learning approach." In: *The Journal of Financial Data Science* 1(3), pp. 9–28.
In this article, the authors use machine learning tools to analyze industry return predictability based on the information in lagged industry returns. Controlling for post-selection inference and multiple testing, they find significant in-sample evidence of industry return predictability. Lagged returns for the financial sector and commodity- and material-producing industries exhibit widespread predictive ability, consistent with the gradual diffusion of information across economically linked industries. Out-of-sample industry return forecasts that incorporate the information in lagged industry returns are economically valuable: Controlling for systematic risk using leading multifactor models from the literature, an industry-rotation portfolio that goes long (short) industries with the highest (lowest) forecasted returns delivers an annualized alpha of over 8%. The industry-rotation portfolio also generates substantial gains during economic downturns, including the Great Recession.

Rapach, D. E. and Zhou, G. (2020). "Time-series and Cross-sectional Stock Return Forecasting: New Machine Learning Methods." In: *Machine Learning for Asset Management: New Developments and Financial Applications*. Ed. by E. Jurczenko. Wiley, pp. 1–33.
Researchers in finance increasingly rely on machine learning techniques to analyze Big Data. This chapter shows how the approach of Han et al., originally designed for forecasting cross-sectional stock returns, can be modified for time-series forecasting of the market excess return. It describes the construction of market excess return forecasts, including the combination elastic net forecast. The chapter reports results for an empirical application centered on forecasting the US market excess return, using a variety of predictor variables from the literature. It outlines the construction of the cross-sectional return forecasts proposed by Han et al. Stock market excess return predictability is typically analyzed in the context of a univariate predictive regression model. Forecast combination reduces forecast "risk" by diversifying across individual forecasts, similarly to diversifying across assets to reduce portfolio risk. Machine learning techniques also provide a means for implementing shrinkage.

Rehman, H.-U., Wan, G., Ullah, A., and Shaukat, B. (2019). "Individual and combination approaches to forecasting hierarchical time series with correlated data: an empirical study." In: *Journal of Management Analytics* 6(3), pp. 231–249.
Hierarchical time series arise in manufacturing and service industries when the products or services have the hierarchical structure, and top-down and bottom-up methods are commonly used to forecast the hierarchical time series. One of the critical factors that affect the performance of the two methods is the correlation between the data series. This study attempts to resolve the problem and shows that the top-down method performs better when data have high positive correlation compared to high negative correlation and combination of forecasting methods may be the best solution when there is no evidence of the correlationship. We conduct the computational experiments using 240 monthly data series from the category of the M3-Competition and test twelve combination methods for the hierarchical data series. The results show that the regression-based, VAR-COV and the Rank-based methods perform better compared to the other methods.

Remlinger, C., Alasseur, C., Briere, M., and Mikael, J. (2022). "Expert Aggregation for Financial Forecasting." In: *SSRN e-Print*.
Machine learning algorithms dedicated to financial time series forecasting have gained a lot of interest over the last few years. One difficulty lies in the choice between several algorithms, as their estimation accuracy may be unstable through time. In this paper, we propose to apply an online aggregation-based forecasting model combining several machine learning techniques to build a portfolio which dynamically adapts itself to market conditions.We apply this aggregation technique to the construction of a long-short-portfolio of individual stocks ranked on their financial characteristics and we demonstrate how aggregation outperforms single algorithms both in terms of performances and of stability.

Reschenhofer, E., Mangat, M. K., Zwatz, C., and Guzmics, S. (2020). "Evaluation of current research on stock return predictability." In: *Journal of Forecasting* 39(2), pp. 334–351.

The results of recent replication studies suggest that false positive findings are a big problem in empirical finance. We contribute to this debate by reviewing a sample of articles dealing with the short-term directional forecasting of the prices of stocks, commodities, and currencies. Screening all relevant articles published in 2016 by one of the 96 journals covered by the Social Sciences Citation Index in the category , Finance, we select only those studies that use easily accessible data of daily or higher frequency. We examine each study in detail, from the selection of the dataset to the interpretation of the results. We also include empirical analyses to illustrate the shortcomings of certain approaches. There are three main findings from our review. First, the number of selected papers is very low, which is surprising even when the strict selection criteria are taken into account. Second, there are hardly any relevant studies that use high-frequency data - despite the hype about financial big data and machine learning. Third, the economic significance of the findings, e.g., their usefulness for trading purposes, is questionable. In general, apparently good forecasting performance does not translate into profitability once realistic transaction costs and the effect of data snooping are taken into account. Other typical problems include unsuitable benchmarks, short evaluation periods, and nonoperational trading strategies.

Risse, M. (2017). "Combining Wavelet Decomposition with Machine Learning to Forecast Gold Returns." In: *SSRN e-Print*.

I combine the discrete wavelet transform with support vector regression to forecast gold-pricedynamics. I investigate the advantages of this approach using a relatively small set of economic and financial predictors. In order to measure model performance, I differentiate between statistical and economic forecast evaluation, where the economic valued-added of forecasts is simulated using a trading strategy. I show that disentangling the predictors with respect to their time and frequency domain leads to improved forecast performance. Results are robust to alternative forecasting approaches. Findings on the relative importances of such wavelet decompositions suggest that the influence of short-term and long-term trends is not stable over the full evaluation period.

Roccazzella, F., Gambetti, P., and Vrins, F. (2022). "Optimal and robust combination of forecasts via constrained optimization and shrinkage." In: *International Journal of Forecasting*.

We introduce various methods that combine forecasts using constrained optimization with penalty. A non-negativity constraint is imposed on the weights, and several penalties are considered, taking the form of a divergence from a reference combination scheme. In contrast with most of the existing approaches, our framework performs forecast selection and combination in one step, allowing for potentially sparse combining schemes. Moreover, by exploiting the analogy between forecasts combination and portfolio optimization, we provide the analytical expression of the optimal penalty strength when penalizing with the L2-divergence from the equally-weighted scheme. An extensive simulation study and two empirical applications allow us to investigate the impact of the divergence function, the reference scheme, and the non-negativity constraint on the predictive performance. Our results suggest that the proposed models outperform those considered in previous studies.

Rossi, B. (2020). "Forecasting in the Presence of Instabilities: How Do We Know Whether Models Predict Well and How to Improve Them." In: *Journal of Economic Literature*.

This article provides guidance on how to evaluate and improve the forecasting ability of models in the presence of instabilities, which are widespread in economic time series. Empirically relevant examples include predicting the financial crisis of 2007-2008, as well as, more broadly, fluctuations in asset prices, exchange rates, output growth and inflation. In the context of unstable environments, I discuss how to assess models' forecasting ability; how to robustify models' estimation; and how to correctly report measures of forecast uncertainty. Importantly, and perhaps surprisingly, breaks in models' parameters are neither necessary nor sufficient to generate time variation in models' forecasting performance: thus, one should not test for breaks in models' parameters, but rather evaluate their forecasting ability in a robust way. In addition, local measures of models' forecasting performance are more appropriate than traditional, average measures.

Roy, R. (2021). "A six-factor asset pricing model: The Japanese evidence." In: *Financial Planning Review* 4(1).

The fundamental research question associated with the asset pricing framework relates to the risk and return relationship in return predictability. We introduce the human capital component to the Fama-French five-factor model and derive an equilibrium six-factor asset pricing model in an intertemporal framework. The study comprises the Japanese monthly time-series dataset for 20 years spanning November 1990 to December 2017. The Generalized method of moments estimation and Gibbons-Ross-Shanken test results confirm that the six-factor model yields better estimates and equally outperforms the Fama-French three-factor, Carhart four-factor, and Fama-French five-factor models in explaining the variation in excess return on Fama-French variant portfolios. The core results and findings hold when we use labor income growth as an alternate measure of human capital in the six-factor asset pricing model.

Rožanec, J., Trajkova, E., Kenda, K., Fortuna, B., and Mladenić, D. (2021). "Explaining Bad Forecasts in Global Time Series Models." In: *Applied Sciences* 11(19), p. 9243.

While increasing empirical evidence suggests that global time series forecasting models can achieve better forecasting performance than local ones, there is a research void regarding when and why the global models fail to provide a good forecast. This paper uses anomaly detection algorithms and explainable artificial intelligence (XAI) to answer when and why a forecast should not be trusted. To address this issue, a dashboard was built to inform the user regarding (i) the relevance of the features for that particular forecast, (ii) which training samples most likely influenced the forecast outcome, (iii) why the forecast is considered an outlier, and (iv) provide a range of counterfactual examples to understand how value changes in the feature vector can lead to a different outcome. Moreover, a modular architecture and a methodology were developed to iteratively remove noisy data instances from the train set, to enhance the overall global time series forecasting model performance. Finally, to test the effectiveness of the proposed approach, it was validated on two publicly available real-world datasets.

Ruan, J., Wu, W., and Luo, J. (2021). "Stock Price Prediction Under Anomalous Circumstances." In: *arXiv e-Print*.

The stock market is volatile and complicated, especially in 2020. Because of a series of global and regional "black swans," such as the COVID-19 pandemic, the U.S. stock market triggered the circuit breaker three times within one week of March 9 to 16, which is unprecedented throughout history. Affected by the whole circumstance, the stock prices of individual corporations also plummeted by rates that were never predicted by any pre-developed forecasting models. It reveals that there was a lack of satisfactory models that could predict the changes in stocks prices when catastrophic, highly unlikely events occur. To fill the void of such models and to help prevent investors from heavy losses during uncertain times, this paper aims to capture the movement pattern of stock prices under anomalous circumstances. First, we detect outliers in sequential stock prices by fitting a standard ARIMA model and identifying the points where predictions deviate significantly from actual values. With the selected data points, we train ARIMA and LSTM models at the single-stock level, industry level, and general market level, respectively. Since the public moods affect the stock market tremendously, a sentiment analysis is also incorporated into the models in the form of sentiment scores, which are converted from comments about specific stocks on Reddit. Based on 100 companies' stock prices in the period of 2016 to 2020, the models achieve an average prediction accuracy of 98% which can be used to optimize existing prediction methodologies.

Ryll, L. and Seidens, S. (2019). "Evaluating the Performance of Machine Learning Algorithms in Financial Market Forecasting: A Comprehensive Survey." In: *arXiv e-Print*.

With increasing competition and pace in the financial markets, robust forecasting methods are becoming more and more valuable to investors. While machine learning algorithms offer a proven way of modeling non-linearities in time series, their advantages against common stochastic models in the domain of financial market prediction are largely based on limited empirical results. The same holds true for determining advantages of certain machine learning architectures against others. This study surveys more than 150 related articles on applying machine learning to financial market forecasting. Based on a comprehensive literature review, we build a table across seven main parameters describing the experiments conducted in these studies. Through listing and classifying different algorithms, we also introduce a simple, standardized syntax for textually representing machine learning algorithms. Based on performance metrics gathered from papers included in the survey, we further conduct rank analyses to assess the comparative performance of different algorithm classes. Our analysis shows that machine learning algorithms tend to outperform most traditional stochastic methods in financial market forecasting. We further find evidence that, on average, recurrent neural networks outperform feed forward neural networks as well as support vector machines which implies the existence of exploitable temporal dependencies in financial time series across multiple asset classes and geographies.

Rytchkov, O. and Zhong, X. (2020). "Information Aggregation and P-Hacking." In: *Management Science* 66(4), pp. 1509–1782.

This paper studies the interplay between information aggregation and p-hacking in the context of predicting stock returns. The standard information-aggregation techniques exacerbate p-hacking by increasing the probability of the type I error. We propose an aggregation technique that is a simple modification of three-pass regression filter/partial least squares regression with an opposite property: the predictability tests applied to the combined predictor become more conservative in the presence of p-hacking. Using simulations, we quantify the advantages of our approach relative to the standard information-aggregation techniques. We also apply our aggregation technique to three sets of return predictors proposed in the literature and find that the forecasting ability of combined predictors in two cases cannot be explained by p-hacking.

Salinas, D., Flunkert, V., and Gasthaus, J. (2020). "DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks." In: *International Journal of Forecasting* 36 (3), pp. 1181–1191.

Probabilistic forecasting, i.e. estimating the probability distribution of a time series' future given its past, is a key enabler for optimizing business processes. In retail businesses, for example, forecasting demand is crucial for having the right inventory available at the right time at the right place. In this paper we propose DeepAR, a methodology for producing accurate probabilistic forecasts, based on training an auto regressive recurrent network model on a large number of related time series. We demonstrate how by applying deep learning techniques to forecasting, one can overcome many of the challenges faced by widely-used classical approaches to the problem. We show through extensive empirical evaluation on several real-world forecasting data sets accuracy improvements of around 15% compared to state-of-the-art methods.

Salisu, A. A. and Tchankam, J. P. (2022). "US Stock return predictability with high dimensional models." In: *Finance Research Letters* 45 (102194), pp. 153–163.

We examine the role of large information sets in the predictability of US stock using a large data set of over 400 predictors covering macro-, financial-, trade- and commodity-related variables over the period of 1960:Q1 to 2018:Q4. We consider 13 alternative models ranging from autoregressive models with no predictors to 5-factor, 60-factor and high dimensional models with over 400 predictors including assumptions of constant and time varying coefficients. We find that models that incorporate large predictors improve US stock return predictability. The outcome particularly favours models involving Dynamic Variable Selection prior with Variational Bayes (VBDV) for density forecast.

Salles, R., Pacitti, E., Bezerra, E., Porto, F., and Ogasawara, E. (2022). "TSPred: A framework for nonstationary time series prediction." In: *Neurocomputing* 467, pp. 197–202.

The nonstationary time series prediction is challenging since it demands knowledge of both data transformation and prediction methods. This paper presents TSPred, a framework for nonstationary time series prediction. It differs from the mainstream frameworks since it establishes a prediction process that seamlessly integrates nonstationary time series transformations with state-of-the-art statistical and machine learning methods. It is made available as an R-package, which provides functions for defining and conducting time series prediction, including data pre(post) processing, decomposition, modeling, prediction, and accuracy assessment. Besides, TSPred enables user-defined methods, which significantly expands the applicability of the framework.

Samuels, J. D. and Sekkel, R. M. (2017). "Model Confidence Sets and forecast combination." In: *International Journal of Forecasting* 33(1), pp. 48–60.

A longstanding finding in the forecasting literature is that averaging the forecasts from a range of models often improves upon forecasts based on a single model, with equal weight averaging working particularly well. This paper analyzes the effects of trimming the set of models prior to averaging. We compare different trimming schemes and propose a new approach based on Model Confidence Sets that takes into account the statistical significance of the out-of-sample forecasting performance. In an empirical application to the forecasting of U.S. macroeconomic indicators, we find significant gains in out-of-sample forecast accuracy from using the proposed trimming method.

Seca, D. (2021). "TimeGym: Debugging for Time Series Modeling in Python." In: *arXiv e-Print*.

We introduce the TimeGym Forecasting Debugging Toolkit, a Python library for testing and debugging time series forecasting pipelines. TimeGym simplifies the testing forecasting pipeline by providing generic tests for forecasting pipelines fresh out of the box. These tests are based on common modeling challenges of time series. Our library enables forecasters to apply a Test-Driven Development approach to forecast modeling, using specified oracles to generate artificial data with noise.

Sharma, P. N., Shmueli, G., Sarstedt, M., Danks, N., and Ray, S. (2020). "Prediction-Oriented Model Selection in Partial Least Squares Path Modeling." In: *Decision Sciences*.

Partial least squares path modeling (PLS-PM) has become popular in various disciplines to model structural relationships among latent variables measured by manifest variables. To fully benefit from the predictive capabilities of PLS-PM, researchers must understand the efficacy of predictive metrics used. In this research, we compare the performance of standard PLS-PM criteria and model selection criteria derived from Information Theory, in terms of selecting the best predictive model among a cohort of competing models. We use Monte Carlo simulation to study this question under various sample sizes, effect sizes, item loadings, and model setups. Specifically, we explore whether, and when, the in-sample measures such as the model selection criteria can substitute for out-of-sample criteria that require a holdout sample. Such a substitution is advantageous when creating a holdout causes considerable loss of statistical and predictive power due to an overall small sample. We find that when the researcher does not have the luxury of a holdout sample, and the goal is selecting correctly specified models with low prediction error, the in-sample model selection criteria, in particular the Bayesian Information Criterion (BIC) and Geweke-Meese Criterion (GM), are useful substitutes for out-of-sample criteria. When a holdout sample is available, the best performing out-of-sample criteria include the root mean squared error (RMSE) and mean absolute deviation (MAD). We recommend against using standard the PLS-PM criteria (R2, Adjusted R2, and Q2), and specifically the out-of-sample mean absolute percentage error (MAPE) for prediction-oriented model selection purposes. Finally, we illustrate the model selection criteria's practical utility using a well-known corporate reputation model.

Shaub, D. (2020). "Fast and accurate yearly time series forecasting with forecast combinations." In: *International Journal of Forecasting* 33(1), pp. 116–120.
It has long been known that combination forecasting strategies produce superior out-of-sample forecasting performances. In the M4 forecasting competition, a very simple forecast combination strategy achieved third place on yearly time series. An analysis of the ensemble model and its component models suggests that the competitive accuracy comes from avoiding poor forecasts, rather than from beating the best individual models. Moreover, the simple ensemble model can be fitted very quickly, can easily scale horizontally with additional CPU cores or a cluster of computers, and can be implemented by users very quickly and easily. This approach might be of particular interest to users who need accurate yearly forecasts without being able to spend significant time, resources, or expertise on tuning models. Users of the R statistical programming language can access this modeling approach using the forecastHybrid package.

Siami-Namini, S., Tavakoli, N., and Namin, A. S. (2019). "A Comparative Analysis of Forecasting Financial Time Series Using ARIMA, LSTM, and BiLSTM." In: *arXiv e-Print*.
Machine and deep learning-based algorithms are the emerging approaches in addressing prediction problems in time series. These techniques have been shown to produce more accurate results than conventional regression-based modeling. It has been reported that artificial Recurrent Neural Networks (RNN) with memory, such as Long Short-Term Memory (LSTM), are superior compared to Autoregressive Integrated Moving Average (ARIMA) with a large margin. The LSTM-based models incorporate additional "gates" for the purpose of memorizing longer sequences of input data. The major question is that whether the gates incorporated in the LSTM architecture already offers a good prediction and whether additional training of data would be necessary to further improve the prediction. Bidirectional LSTMs (BiLSTMs) enable additional training by traversing the input data twice (i.e., 1) left-to-right, and 2) right-to-left). The research question of interest is then whether BiLSTM, with additional training capability, outperforms regular unidirectional LSTM. This paper reports a behavioral analysis and comparison of BiLSTM and LSTM models. The objective is to explore to what extend additional layers of training of data would be beneficial to tune the involved parameters. The results show that additional training of data and thus BiLSTM-based modeling offers better predictions than regular LSTM-based models. More specifically, it was observed that BiLSTM models provide better predictions compared to ARIMA and LSTM models. It was also observed that BiLSTM models reach the equilibrium much slower than LSTM-based models.

Siebert, J., Gross, J., and Schroth, C. (2021). "A systematic review of Python packages for time series analysis." In: *Engineering Proceedings* 5(1) (22).
This paper presents a systematic review of Python packages focused on time series analysis. The objective is first to provide an overview of the different time series analysis tasks and preprocessing methods implemented, but also to give an overview of the development characteristics of the packages (e.g., dependencies, community size, etc.). This review is based on a search of literature databases as well as GitHub repositories. After the filtering process, 40 packages were analyzed. We classified the packages according to the analysis tasks implemented, the methods related to data preparation, and the means to evaluate the results produced (methods and access to evaluation data). We also reviewed the licenses, the packages community size, and the dependencies used. Among other things, our results show that forecasting is by far the most implemented task, that half of the packages provide access to real datasets or allow generating synthetic data, and that many packages depend on a few libraries (the most used ones being numpy, scipy and pandas). One of the lessons learned from this review is that the process of finding a given implementation is not inherently simple, and we hope that this review can help practitioners and researchers navigate the space of Python packages dedicated to time series analysis.

Siliverstovs, B. and Wochner, D. (2021). "State-Dependent Evaluation of Predictive Ability." In: *Journal of Forecasting* 40(3), pp. 547–574.

This study systematically broadens the relevance of possible model performance asymmetries across business cycles in the spirit of the recent state-dependent forecast evaluation literature (e.g. Chauvet and Potter, 2013) to hundreds of macroeconomic indicators and deepens the forecast evaluation of the recent factor model literature on hundreds of target variables (e.g. Stock and Watson, 2012b) in a state-dependent manner. Our results are consistent with both strands of the literature and generalize the former to over 200 macroeconomic indicators and differentiate the latter across three levels of temporal granularity: We document systematic model performance differences in both absolute and relative terms across business cycles (longitudinal) as well as across variable groups (cross-sectional) and find these performance differences to be robust across several alternative specifications. The cross-sectional prevalence and robustness of state-dependency shown in this article encourages economic forecasters to complement model performance assessments with a state-dependent evaluation of predictive ability.

Smith, S. C., Bulkley, G., and Leslie, D. S. (2020). "Equity Premium Forecasts with an Unknown Number of Structural Breaks." In: *Journal of Financial Econometrics* 18(1), pp. 59–94.
Estimation of models with structural breaks usually assumes a pre-specified number of breaks. Previous models which do allow an endogenously determined number of breaks require a simple structural model, and rarely allow for information transfer across the break. We introduce a methodology that allows the number of breaks to be determined endogenously and including an economically motivated model of transition regimes between each break. We demonstrate the usefulness of our approach for forecasts of the equity premium. We find the demonstrated success of the historical average can be improved upon by an economic model with theory informed priors estimated using our methodology.

Smith, S. C. and Timmermann, A. (2021). "Break Risk." In: *The Review of Financial Studies* 34(4), pp. 2045–2100.
We develop a new approach to modeling and predicting stock returns in the presence of breaks that simultaneously affect a large cross-section of stocks. Exploiting information in the cross-section enables us to detect breaks in return prediction models with little delay and to generate out-of-sample return forecasts that are significantly more accurate than those from existing approaches. To identify the economic sources of breaks, we explore the asset pricing restrictions implied by a present value model which links breaks in return predictability to breaks in the cash flow growth and discount rate processes.

Smyl, S. (2020). "A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting." In: *International Journal of Forecasting* 36(1) (1), pp. 75–85.
This paper presents the winning submission of the M4 forecasting competition. The submission utilizes a dynamic computational graph neural network system that enables a standard exponential smoothing model to be mixed with advanced long short term memory networks into a common framework. The result is a hybrid and hierarchical forecasting method.

Son, B. and Lee, J. (2022). "Graph-based multi-factor asset pricing model." In: *Finance Research Letters* 44 (102032).
We propose a latent multi-factor asset pricing model that estimates risk exposure based on firm characteristics and connectivity between assets. To handle connected high-dimensional characteristics, we adopted a graph convolutional network while estimating the connectivity between assets from the correlation of asset returns. Unlike recent literature involving the deep-learning-based latent factor model, we propose a forward stagewise additive factor modeling architecture that constructs latent factors sequentially to maintain the previous stage's factors. Our empirical results on individual U.S. equities show that the proposed graph factor model outperforms other benchmark models in terms of explanatory power and the Sharpe ratio of the factor tangency portfolio.

Spiliotis, E., Abolghasemi, M., Hyndman, R. J., Petropoulos, F., and Assimakopoulos, V. (2021). "Hierarchical forecast reconciliation with machine learning." In: *Applied Soft Computing* 112, p. 107756.
Over the last 15 years, studies on hierarchical forecasting have moved away from single-level approaches towards proposing linear combination approaches across multiple levels of the hierarchy. Such combinations offer coherent reconciled forecasts, improved forecasting performance and aligned decision-making. This paper proposes a novel hierarchical forecasting approach based on machine learning. The proposed method allows for non-linear combinations of the base forecasts, thus being more general than linear approaches. We structurally combine the objectives of improved post-sample empirical forecasting accuracy and coherence. Due to its non-linear nature, our approach selectively combines the base forecasts in a direct and automated way without requiring that the complete information must be used for producing reconciled forecasts for each series and level. The proposed method is evaluated both in terms of accuracy and bias using two different data sets coming from the tourism and retail industries. Our results suggest that the proposed method gives superior point forecasts than existing approaches, especially when the series comprising the hierarchy are not characterized by the same patterns.

Spiliotis, E., Nikolopoulos, K., and Assimakopoulos, V. (2019). "Tales from tails: On the empirical distributions of forecasting errors and their implication to risk." In: *International Journal of Forecasting* 35(2), pp. 687–698.
Abstract When evaluating the performances of time series extrapolation methods, both researchers and practitioners typically focus on the average or median performance according to some specific error metric, such as the absolute error or the absolute percentage error. However, from a risk-assessment point of view, it is far more important to evaluate the distributions of such errors, and especially their tails. For instance, a lack of normality and symmetry in error distributions can have significant implications for decision making, such as in stock control. Moreover, frequently these distributions can only be constructed empirically, as they may be the result of a computationally-intensive non-parametric approach, such as an artificial neural network. This study proposes an approach for evaluating the empirical distributions of forecasting methods and uses it to assess eleven popular time series extrapolation approaches across two different datasets (M3 and ForeDeCk). The results highlight some very interesting tales from the tails.

Stauskas, O. and Westerlund, J. (2022). "Tests of Equal Forecasting Accuracy for Nested Models with Estimated CCE Factors." In: *Journal of Business & Economic Statistics*, pp. 1–14.
In this article, we propose new tests of equal predictive ability between nested models when factor-augmented regressions are used to forecast. In contrast to the previous literature, the unknown factors are not estimated by principal components but by the common correlated effects (CCE) approach, which employs cross-sectional averages of blocks of variables. This makes for easy interpretation of the estimated factors, and the resulting tests are easy to implement and they account for the block structure of the data. Assuming that the number of averages is larger than the true number of factors, we establish the limiting distributions of the new tests as the number of time periods and the number of variables within each block jointly go to infinity. The main finding is that the limiting distributions do not depend on the number of factors but only on the number of averages, which is known. The important practical implication of this finding is that one does not need to estimate the number of factors consistently in order to apply our tests.

Stein, T. (2021). "Out-of-Sample Equity Premium Prediction: Combination Forecasts with Frequency-Decomposed Variables." In: *2nd Frontiers of Factor Investing Conference*.
Technical trading rules are widely used by practitioners to forecast the U.S. equity premium. I decompose technical indicators into components with frequency specific information, showing that all the predictive power comes from periodicities between 16 to 64 months, without any evidence of predictability outside of this frequency band. An investor who only forecasts with these medium-frequency components generates both statistically and economically sizable gains compared to the historical mean and the original technical indicators. The out-of-sample R2 is significant for each of the 14 adjusted indicators in the sample. A mean-variance investor who combines individual forecasts from medium-frequency components generates a sizable utility gain of more than 350 basis points relative to the historical mean for the forecasting period from January 1966 to December 2017. This is almost twice as large as utility gains from the historical mean and more than 200 basis points larger than for combination forecasts with unadjusted technical indicators. I show that the substantial gains mainly result from an improved forecasting ability of mediumfrequency components during recessions.

Stivers, A. (2018). "Equity premium predictions with many predictors: A risk-based explanation of the size and value factors." In: *Journal of Empirical Finance* 45, pp. 126–140.
This paper investigates whether a direct mechanism can be found that demonstrates that the size and value factors of Fama and French (1993) are indeed ICAPM factors, as some have suggested. The results endorse this hypothesis: small size and value portfolios reflect changes in future investment opportunities. To test the hypothesis, the paper forecasts the equity premium using disaggregated portfolio returns with a partial least squares (PLS) regression approach. PLS is chosen as it is particularly suited to condense a large set of portfolios into a single index. The "index" portfolio obtained from the forecast performs well out of sample and hedges against future market risk, in addition to explaining future market returns in sample. Thus, the index portfolio may be viewed as an additional risk factor in the form of a Merton (1973) state variable. The index places larger weights on small size and value portfolios. This also provides a possible explanation for why equal-weighted portfolios typically perform better out of sample compared to factors implied by traditional mean-variance approaches and asset pricing models.

Stoyanov, S. V. and Fabozzi, F. A. (2021). "Dynamics of Equity Factor Returns and Asset Pricing." In: *Journal of Financial Econometrics*.
In empirical equity asset pricing, the stochastic discount factor (SDF) is implicitly modeled as a linear function of equity factors and is influenced by the empirical properties of the factor returns. We investigate the pricing

error introduced by a misspecified SDF which ignores each of the following established empirical phenomena: autocorrelation, dynamics of covariances, dynamics of correlations, and heavy tails for the conditional factor return distribution. We consider near-linear SDFs and nonlinear specifications characterized by a high degree of risk aversion. We find that assuming constant covariances or constant correlations can significantly overprice certain equity portfolios at all risk-aversion levels and that ignoring fat tails can lead to large pricing errors for some derivative assets for highly nonlinear SDFs.

Suhonen, A., Lennkh, M., and Perez, F. (2017). "Quantifying Backtest Overfitting in Alternative Beta Strategies." In: *The Journal of Portfolio Management* 43 (2), pp. 90–104.

The authors investigate the biases in the backtested performance of ”alternative beta”strategies using a unique sample of 215 trading strategies developed and promoted by global investment banks. Their results lend support to the cautions in the recent literature regarding backtest overfitting and lack of robustness in trading strategy performance during the ”live”period (out of sample). The authors report a median 73 percent deterioration in Sharpe ratios between backtested and live performance periods for the strategies, and they establish a link between performance deterioration and strategy complexity, with the realized reduction in live versus backtested Sharpe ratios of the most complex strategies exceeding those of the simplest ones by over 30 percentage points. The robustness of strategy exposure to risk factors varies between asset classes and strategies; it appears reasonable in equity volatility and FX carry strategies but quite weak in the equity value strategy in particular.

Svensson, M. (2018). "An Evaluation of Methods for Combining Univariate Time Series Forecasts." MA thesis. Lund University.

This thesis presents and evaluates nineteen methods for combining up to eleven automated univariate forecasts. The evaluation is made by applying the methods on a dataset containing more than 1000 monthly time series. The accuracy of one period ahead forecasts is analyzed. Almost 3.2 million forecasts are evaluated in the study. Methods that are using past forecasts to optimally produce a combined forecast are included, along with methods that do not require this information. A pre-screening procedure to get rid of the poorest performing forecasting methods before the remaining ones are combined is evaluated. The results confirm that it is possible to achieve a superior forecast accuracy by combining forecasts. The best methods that utilize past forecasts tend to outperform the best methods that are not considering this data. Including a pre-screening procedure to remove inferior forecasts before combining forecasts from the top five ranked methods seems to increase the forecast accuracy. The pre-screening procedure consists of ranking the automated univariate forecasting methods using an independent, but relevant, dataset. The four best performing methods utilize the pre-screening procedure together with past forecasts to optimally combine forecasts. The best method computes the historical mean squared error of each individual method and weights them accordingly. Demand for automated procedures is growing as the size of datasets increases within organizations. Forecasting from a large set of time series is an activity that can take advantage of automated procedures. However, choosing which forecasting method to use is often problematic. One way of solving this is by combining multiple forecasts into a single forecast.

Tadayon, M. and Iwashita, Y. (2020). "Comprehensive Analysis of Time Series Forecasting Using Neural Networks." In: *arXiv e-Print*.

Time series forecasting has gained lots of attention recently; this is because many real-world phenomena can be modeled as time series. The massive volume of data and recent advancements in the processing power of the computers enable researchers to develop more sophisticated machine learning algorithms such as neural networks to forecast the time series data. In this paper, we propose various neural network architectures to forecast the time series data using the dynamic measurements; moreover, we introduce various architectures on how to combine static and dynamic measurements for forecasting. We also investigate the importance of performing techniques such as anomaly detection and clustering on forecasting accuracy. Our results indicate that clustering can improve the overall prediction time as well as improve the forecasting performance of the neural network. Furthermore, we show that feature-based clustering can outperform the distance-based clustering in terms of speed and efficiency. Finally, our results indicate that adding more predictors to forecast the target variable will not necessarily improve the forecasting accuracy.

Taggart, R. J. (2021). "Evaluation of point forecasts for extreme events using consistent scoring functions." In: *arXiv e-Print*.

We present a method for comparing point forecasts in a region of interest, such as the tails or centre of a variable's range. This method cannot be hedged, in contrast to conditionally selecting events to evaluate and then using a scoring function that would have been consistent (or proper) prior to event selection. Our method also gives decompositions of scoring functions that are consistent for the mean or a particular quantile or expectile. Each member of each decomposition is itself a consistent scoring function that emphasises performance over a selected region of the variable's range. The score of each member of the decomposition has a natural interpretation rooted in optimal decision theory. It is the weighted average of economic regret over user decision thresholds, where the weight emphasises those decision thresholds in the corresponding region of interest.

Taillardat, M., Fougeres, A.-L., Naveau, P., and de Fondeville, R. (2022). "Extreme events evaluation using CRPS distributions." In: *arXiv e-Print*.

Verification of ensemble forecasts for extreme events remains a challenging question. The general public as well as the media naturely pay particular attention on extreme events and conclude about the global predictive performance of ensembles, which are often unskillful when they are needed. Ashing classical verification tools to focus on such events can lead to unexpected behaviors. To square up these effects, thresholded and weighted scoring rules have been developed. Most of them use derivations of the Continuous Ranked Probability Score (CRPS). However, some properties of the CRPS for extreme events generate undesirable effects on the quality of verification. Using theoretical arguments and simulation examples, we illustrate some pitfalls of conventional verification tools and propose a different direction to assess ensemble forecasts using extreme value theory, considering proper scores as random variables.

Talagala, T. S., Li, F., and Kang, Y. (2021). "FFORMPP: Feature-based forecast model performance prediction." In: *arXiv e-Print*.

This paper introduces a novel meta-learning algorithm for time series forecast model performance prediction. We model the forecast error as a function of time series features calculated from the historical time series with an efficient Bayesian multivariate surface regression approach. The minimum predicted forecast error is then used to identify an individual model or a combination of models to produce the final forecasts. It is well-known that the performance of most meta-learning models depends on the representativeness of the reference dataset used for training. In such circumstances, we augment the reference dataset with a feature-based time series simulation approach, namely GRATIS, in generating a rich and representative time series collection. The proposed framework is tested using the M4 competition data and is compared against commonly used forecasting approaches. Our approach provides provides comparable performances to other model selection/combination approaches but at lower computational cost and higher degree of interpretability, which is important for supporting decisions. We also provide useful insights regarding which forecasting models are expected to work better for particular types of time series, how the meta-learners work and how the forecasting performances are affected by various factors.

Talagala, T. S., Li, F., and Kang, Y. (2022). "FFORMPP: Feature-based forecast model performance prediction." In: *International Journal of Forecasting*.

This paper introduces a novel meta-learning algorithm for time series forecast model performance prediction. We model the forecast error as a function of time series features calculated from historical time series with an efficient Bayesian multivariate surface regression approach. The minimum predicted forecast error is then used to identify an individual model or a combination of models to produce the final forecasts. It is well known that the performance of most meta-learning models depends on the representativeness of the reference dataset used for training. In such circumstances, we augment the reference dataset with a feature-based time series simulation approach, namely GRATIS, to generate a rich and representative time series collection. The proposed framework is tested using the M4 competition data and is compared against commonly used forecasting approaches. Our approach provides comparable performance to other model selection and combination approaches but at a lower computational cost and a higher degree of interpretability, which is important for supporting decisions. We also provide useful insights regarding which forecasting models are expected to work better for particular types of time series, the intrinsic mechanisms of the meta-learners, and how the forecasting performance is affected by various factors.

Tang, X., Hu, F., and Wang, P. (2018). "Out-of-sample equity premium prediction: A scenario analysis approach." In: *Journal of Forecasting* 37(5), pp. 604–626.

We propose two methods of equity premium prediction with single and multiple predictors respectively and evaluate their out-of-sample performance using US stock data with 15 popular predictors for equity premium prediction. The first method defines three scenarios in terms of the expected returns under the scenarios and assumes a Markov chain governing the occurrence of the scenarios over time. It employs predictive quantile regressions of excess return on a predictor for three quantiles to estimate the occurrence of the scenarios over an in-sample period and the transition probabilities of the Markov chain, predicts the expected returns under the scenarios, and generates an equity premium forecast by combining the predicted expected returns under three

scenarios with the estimated transition probabilities. The second method generates an equity premium forecast by combining the individual forecasts from the first method across all predictors. For most of predictors, the first method outperforms the benchmark method of historical average and the traditional predictive linear regression with a single predictor both statistically and economically, and the second method consistently performs better than several competing methods used in the literature. The performance of our methods is further examined under different scenarios and economic conditions, and is robust for two different out-of-sample periods and specifications of the scenarios.

Taylor, J. W. and Taylor, K. S. (2021). "Combining probabilistic forecasts of COVID-19 mortality in the United States." In: *European Journal of Operational Research*.

The COVID-19 pandemic has placed forecasting models at the forefront of health policy making. Predictions of mortality, cases and hospitalisations help governments meet planning and resource allocation challenges. In this paper, we consider the weekly forecasting of the cumulative mortality due to COVID-19 at the national and state level in the U.S. Optimal decision-making requires a forecast of a probability distribution, rather than just a single point forecast. Interval forecasts are also important, as they can support decision making and provide situational awareness. We consider the case where probabilistic forecasts have been provided by multiple forecasting teams, and we combine the forecasts to extract the wisdom of the crowd. We use a dataset that has been made publicly available from the COVID-19 Forecast Hub. A notable feature of the dataset is that the availability of forecasts from participating teams varies greatly across the 40 weeks in our study. We evaluate the accuracy of combining methods that have been previously proposed for interval forecasts and predictions of probability distributions. These include the use of the simple average, the median, and trimming methods. In addition, we propose several new weighted combining methods. Our results show that, although the median was very useful for the early weeks of the pandemic, the simple average was preferable thereafter, and that, as a history of forecast accuracy accumulates, the best results can be produced by a weighted combining method that uses weights that are inversely proportional to the historical accuracy of the individual forecasting teams.

Theodosiou, F. and Kourentzes, N. (2021). "Forecasting with Deep Temporal Hierarchies." In: *SSRN e-Print*.

In time series analysis and forecasting, the identification of an appropriate model remains a challenging task. Model misspecification can lead to erroneous forecasts and insights. The use of multiple views of the same time series by constructing temporally aggregate levels has been proposed as a way to overcome the model specification and selection uncertainty, with ample empirical evidence of forecast accuracy gains. Temporal Hierarchies is the most popular approach to achieve this, which itself is based on research in hierarchical forecasting. Although there has been substantial progress in this literature, the vast majority of methods rely on a restricted linear combination of different model outputs across the hierarchy. We investigate the use of deep learning to augment temporal hierarchies, relaxing the classical restrictions. Specifically, we look at deep learning for the generation of all the base forecasts, the hierarchical reconciliation, and an end-to-end method that embeds all steps in a single neural network. We inspect the performance of the proposed methods when applied to individual time series, or with global training across complete sets of series. We further investigate the requirements in terms of series set size, illustrating the conditions where deep learning temporal hierarchies outperform conventional temporal hierarchies.

Thomson, M. E., Pollock, A. C., Onkal, D., and Gonul, M. S. (2019). "Combining forecasts: Performance and coherence." In: *International Journal of Forecasting* 35(2), pp. 474–484.

There is general agreement in many forecasting contexts that combining individual predictions leads to better final forecasts. However, the relative error reduction in a combined forecast depends upon the extent to which the component forecasts contain unique/independent information. Unfortunately, obtaining independent predictions is difficult in many situations, as these forecasts may be based on similar statistical models and/or overlapping information. The current study addresses this problem by incorporating a measure of coherence into an analytic evaluation framework so that the degree of independence between sets of forecasts can be identified easily. The framework also decomposes the performance and coherence measures in order to illustrate the underlying aspects that are responsible for error reduction. The framework is demonstrated using UK retail prices index inflation forecasts for the period 1998-2014, and implications for forecast users are discussed.

Thorarinsdottir, T. L. (2021). "Forecast evaluation." In: *CUSO winter school*.

Presentation in 3 parts:

1) Part One: `https://statistique.cuso.ch/fileadmin/statistique/user_upload/TLT_part1.pdf`
2) Part Two `https://statistique.cuso.ch/fileadmin/statistique/user_upload/TLT_part2..final.pdf`

3) Part Three https://statistique.cuso.ch/fileadmin/statistique/user_upload/TLT_part3.Final.pdf

.

Tilly, S., Ebner, M., and Livan, G. (2021). "Macroeconomic forecasting through news, emotions and narrative." In: *Expert Systems with Applications* 175, p. 114760.
This study proposes a new method of incorporating emotions from newspaper articles into macroeconomic forecasts, attempting to forecast industrial production and consumer prices leveraging narrative and sentiment from global newspapers. For the most part, existing research includes positive and negative tone only to improve macroeconomic forecasts, focusing predominantly on large economies such as the US. These works use mainly anglophone sources of narrative, thus not capturing the entire complexity of the multitude of emotions contained in global news articles. This study expands the existing body of research by incorporating a wide array of emotions from newspapers around the world - extracted from the Global Database of Events, Language and Tone (GDELT) - into macroeconomic forecasts. We present a thematic data filtering methodology based on a bi-directional long short term memory neural network (Bi-LSTM) for extracting emotion scores from GDELT and demonstrate its effectiveness by comparing results for filtered and unfiltered data. We model industrial production and consumer prices across a diverse range of economies using an autoregressive framework, and find that including emotions from global newspapers significantly improves forecasts compared to three autoregressive benchmark models. We complement our forecasts with an interpretability analysis on distinct groups of emotions and find that emotions associated with happiness and anger have the strongest predictive power for the variables we predict.

Tilly, S. and Livan, G. (2021). "Macroeconomic forecasting with statistically validated knowledge graphs." In: *Expert Systems with Applications* 186, p. 115765.
This study leverages narrative from global newspapers to construct theme-based knowledge graphs about world events, demonstrating that features extracted from such graphs improve forecasts of industrial production in three large economies compared to a number of benchmarks. Our analysis relies on a filtering methodology that extracts ”backbones” of statistically significant edges from large graph data sets. We find that changes in the eigenvector centrality of nodes in such backbones capture shifts in relative importance between different themes significantly better than graph similarity measures. We supplement our results with an interpretability analysis, showing that the theme categories ”disease” and ”economic” have the strongest predictive power during the time period that we consider. Our work serves as a blueprint for the construction of parsimonious - yet informative - theme-based knowledge graphs to monitor in real time the evolution of relevant phenomena in socio-economic systems.

Timmermann, A. (2018). "Forecasting methods in finance." In: *Annual Review of Financial Economics* 10(1), pp. 449–470.
Our review highlights some of the key challenges in financial forecasting problems and opportunities arising from the unique features of financial data. We analyze the difficulty of establishing predictability in an environment with a low signal-to-noise ratio, persistent predictors, and instability in predictive relations arising from competitive pressures and investors' learning. We discuss approaches for forecasting the mean, variance, and probability distribution of asset returns. Finally, we discuss how to evaluate financial forecasts while accounting for the possibility that numerous forecasting models may have been considered, leading to concerns of data mining.

Trucíos, C., Mazzeu, J. H. G., Hallin, M., Hotta, L. K., Pereira, P. L. V., and Zevallos, M. (2021). "Forecasting Conditional Covariance Matrices in High-Dimensional Time Series: a General Dynamic Factor Approach." In: *Journal of Business & Economic Statistics*, pp. 1–35.
Based on a General Dynamic Factor Model with infinite-dimensional factor space and MGARCH volatility models, we develop new estimation and forecasting procedures for conditional covariance matrices in high-dimensional time series. The finite-sample performance of our approach is evaluated via Monte Carlo experiments and outperforms most alternative methods. This new approach is also used to construct minimum one-step-ahead variance portfolios for a high-dimensional panel of assets. The results are shown to match the results of recent proposals by Engle et al. (2019) and De Nard et al. (2021) and achieve better out-of-sample portfolio performance than alternative procedures proposed in the literature.

Tunaru, D., Fabozzi, F. A., and Fabozzi, F. J. (2021). "Testing the Forecasting Ability of Multi-Factor Models on Non-US Interbank Rates." In: *The Journal of Fixed Income* 31(2).

This article examines the forecasting performance of continuous-time multi-factor models, in comparison with other parsimonious models, for the term structure of interbank rates in the UK, Europe, and Japan. The article employs two general dynamic frameworks with different factor structures: the generalized Chan-Karolyi-Longstaff-Sanders family of models and the arbitrage-free dynamic Nelson-Siegel family of models. Applying a battery of accuracy measures and a range of formal tests of forecasting superiority, this research provides evidence that extended multi-factor models demonstrate good out-of-sample forecasting performance for the short segment of the yield curve. However, for the euro and in part for the yen, random walk forecasts consistently pass various tests, indicating a higher level of market efficiency compared to the pound sterling interbank market.

Vaiciukynas, E., Danenas, P., Kontrimas, V., and Butleris, R. (2022). "Two-Step Meta-Learning for Time-Series Forecasting Ensemble." In: *IEEE Access* 9, pp. 62687–62696.

Amounts of historical data collected increase and business intelligence applicability with automatic forecasting of time series are in high demand. While no single time series modeling method is universal to all types of dynamics, forecasting using an ensemble of several methods is often seen as a compromise. Instead of fixing ensemble diversity and size, we propose to predict these aspects adaptively using meta-learning. Meta-learning here considers two separate random forest regression models, built on 390 time-series features, to rank 22 univariate forecasting methods and recommend ensemble size. The forecasting ensemble is consequently formed from methods ranked as the best, and forecasts are pooled using either simple or weighted average (with a weight corresponding to reciprocal rank). The proposed approach was tested on 12561 micro-economic time-series (expanded to 38633 for various forecasting horizons) of M4 competition where meta-learning outperformed Theta and Comb benchmarks by relative forecasting errors for all data types and horizons. Best overall results were achieved by weighted pooling with a symmetric mean absolute percentage error of 9.21% versus 11.05% obtained using the Theta method.

van Dijk, D. and Franses, P. H. (2019). "Combining expert-adjusted forecasts." In: *Journal of Forecasting* 38(5), pp. 415–421.

It is well known that a combination of model-based forecasts can improve upon each of the individual constituent forecasts. Most forecasts available in practice are, however, not purely based on econometric models but entail adjustments, where experts with domain-specific knowledge modify the original model forecasts. There is much evidence that expert-adjusted forecasts do not necessarily improve the pure model-based forecasts. In this paper we show, however, that combined expert-adjusted model forecasts can improve on combined model forecasts, in the case when the individual expert-adjusted forecasts are not better than their associated model-based forecasts. We discuss various implications of this finding.

Vincent, K., Hsu, Y.-C., and Lin, H.-W. (2020). "Investment styles and the multiple testing of cross-sectional stock return predictability." In: *Journal of Financial Markets*.

The scheme of simultaneously testing many profitable strategies may conceal the hazard of data-snooping bias. However, certain portfolio returns are also more likely to exhibit codependency because of their same investment styles. Aiming at the phenomena of stock return anomalies, we consider two multiple testing approaches: one ignores the classification of portfolios and the other utilizes such information. The results based on grouped multiple testing suggest that the implied adjusted critical values for t-statistics may vary across investment styles, and several statistically significant portfolios may be unidentified under the pooled setup.

Viswanathan, T. and Stephen, M. (2020). "Does Machine Learning Algorithms Improve Forecasting Accuracy? Predicting Stock Market Index Using Ensemble Model." In: *Advances in Distributed Computing and Machine Learning*. Springer Singapore, pp. 511–519.

Forecasting market performance and understanding the mechanism of price discovery is inherent to develop trading strategies. This paper examines the predictive power of machine learning algorithms in forecasting stock index. The study applies various machine learning algorithms and suggests the best model for forecasting stock index. We have developed an ensemble model to predict the daily closing index of Nifty 50 based on open, high, low and previous day's close. The ensemble model includes a mix of simple linear regression, gradient boosted tree, decision tree and random forest. The parameters of the models are tested for its accuracy train-test split under supervised learning. The predicting accuracy of machine learning algorithms is further refined using cross-validation techniques that include leaving one out cross-validation and k-fold cross-validation. We found that the ensemble model provides an accurate forecast of the stock market index for the short term. The outcome of the study would facilitate the investors and portfolio managers to use the appropriate model for forecasting and take an informed decision by considering the nature of stock market volatility.

Vovk, V. and Wang, R. (2021). "E-values: Calibration, combination, and applications." In: *Annals of Statistics* 49(3), pp. 1736–1753.

Multiple testing of a single hypothesis and testing multiple hypotheses are usually done in terms of p-values. In this paper we replace p-values with their natural competitor, e-values, which are closely related to betting, Bayes factors, and likelihood ratios. We demonstrate that e-values are often mathematically more tractable; in particular, in multiple testing of a single hypothesis, e-values can be merged simply by averaging them. This allows us to develop efficient procedures using e-values for testing multiple hypotheses.

Wang, H., Ahluwalia, H. S., Aliaga-Diaz, R. A., and Davis, J. H. (2021a). "The Best of Both Worlds: Forecasting US Equity Market Returns Using a Hybrid Machine Learning Time Series Approach." In: *The Journal of Financial Data Science* 3(2), pp. 9–20.

Predicting long term equity market returns is of great importance for investors to strategically allocate their assets. The authors explore machine learning (ML) methods to forecast 10 year ahead US stock returns and compare the results with the traditional Shiller regression based forecasts more commonly used in the asset-management industry. The authors find that ML techniques can only modestly improve the forecast accuracy of a traditional Shiller cyclically adjusted price to earnings ratio model, and they actually result in worse performance than the vector autoregressive model (VAR) based two step approach. The authors then implement this approach with ML techniques and allow for unspecified nonlinear relationships (a hybrid ML VAR approach). They find about 50% improvement in real-time forecast accuracy for 10 year annualized US stock returns.

Wang, R. and Ramdas, A. (2020). "False discovery rate control with e-values." In: *arXiv e-Print*.

E-values have gained recent attention as potential alternatives to p-values as measures of uncertainty, significance and evidence. In brief, e-values are random variables with expectation at most one under the null; examples include betting scores, inverse Bayes factors, likelihood ratios and stopped supermartingales. We design a natural analog of the Benjamini-Hochberg (BH) procedure for false discovery control (FDR) control that utilizes e-values (e-BH) and compare it with the standard procedure for p-values. One of our central results is that, unlike the usual BH procedure, the e-BH procedure controls the FDR at the desired level—*with no correction*—for any dependence structure between the e-values. We show that the e-BH procedure includes the BH procedure as a special case through calibration between p-values and e-values. Several illustrative examples and results of independent interest are provided.

Wang, Y., Hao, X., and Wu, C. (2021b). "Forecasting stock returns: A time-dependent weighted least squares approach." In: *Journal of Financial Markets* 53 (100568), p. 100568.

We improve the performance of stock return forecasts using predictive regressions with ordinary least squares (OLS) estimates weighted by a class of time-dependent functions (TWLS). To address the structural breaks in predictive relationships, these functions assign heavier weights to more recent observations. We find return predictability that is statistically and economically significant using a forecast combination of univariate TWLS models. TWLS estimates lead to much stronger return predictability than OLS estimates. The forecast improvement from TWLS is also found when forecasting characteristic portfolio returns and when using newly proposed predictor variables. These findings survive a series of robustness checks.

Wang, Y., Smola, A., Maddix, D., Gasthaus, J., Foster, D., and Januschowski, T. (2019). "Deep Factors for Forecasting." In: *Proceedings of Machine Learning Research* 97, pp. 6607–6617.

Producing probabilistic forecasts for large collections of similar and/or dependent time series is a practically highly relevant, yet challenging task. Classical time series models fail to capture complex patterns in the data and multivariate techniques struggle to scale to large problem sizes, but their reliance on strong structural assumptions makes them data-efficient and allows them to provide estimates of uncertainty. The converse is true for models based on deep neural networks, which can learn complex patterns and dependencies given enough data. In this paper, we propose a hybrid model that incorporates the benefits of both approaches. Our new method is data-driven and scalable via a latent, global, deep component. It also handles uncertainty through a local classical model. We provide both theoretical and empirical evidence for the soundness of our approach through a necessary and sufficient decomposition of exchangeable time series into a global and a local part and extensive experiments. Our experiments demonstrate the advantages of our model both in term of data efficiency and computational complexity.

Weigand, A. (2019). "Machine learning in empirical asset pricing." In: *Financial Markets and Portfolio Management* 33, pp. 93–104.

The tremendous speedup in computing in recent years, the low data storage costs of today, the availability of data as well as the broad range of free open-source software, have created a renaissance in the application of machine learning techniques in science. However, this new wave of research is not limited to computer science or software engineering anymore. Among others, machine learning tools are now used in financial problem settings as well. Therefore, this paper mentions a specific definition of machine learning in an asset pricing context and elaborates on the usefulness of machine learning in this context. Most importantly, the literature review gives the reader a theoretical overview of the most recent academic studies in empirical asset pricing that employ machine learning techniques. Overall, the paper concludes that machine learning can offer benefits for future research. However, researchers should be critical about these methodologies as machine learning has its pitfalls and is relatively new to asset pricing.

Weiss, C. E., Raviv, E., and Roetzer, G. (2018). "Forecast Combinations in R using the ForecastComb Package." In: *The R Journal* 10(2), pp. 262–281.
This paper introduces the R package ForecastComb. The aim is to provide researchers and practitioners with a comprehensive implementation of the most common ways in which forecasts can be combined. The package in its current version covers 15 popular estimation methods for creating a combined forecasts - including simple methods, regression-based methods, and eigenvector-based methods. It also includes useful tools to deal with common challenges of forecast combination (e.g., missing values in component forecasts, or multicollinearity), and to rationalize and visualize the combination results.

Wellens, A. P., Udenio, M., and Boute, R. N. (2022). "Transfer learning for hierarchical forecasting: Reducing computational efforts of M5 winning methods." In: *International Journal of Forecasting*.
The winning machine learning methods of the M5 Accuracy competition demonstrated high levels of forecast accuracy compared to the top-performing benchmarks in the history of the M-competitions. Yet, large-scale adoption is hampered due to the significant computational requirements to model, tune, and train these state-of-the-art algorithms. To overcome this major issue, we discuss the potential of transfer learning (TL) to reduce the computational effort in hierarchical forecasting and provide a proof of concept that TL can be applied on M5 top-performing methods. We demonstrate our easy-to-use TL framework on the recursive store-level LightGBM models of the M5 winning method and attain similar levels of forecast accuracy with roughly 25% less training time. Our findings provide evidence for a novel application of TL to facilitate the practical applicability of the M5 winning methods in large-scale settings with hierarchically structured data.

Wen, D., He, M., Zhang, Y., and Wang, Y. (2022). "Forecasting realized volatility of Chinese stock market: A simple but efficient truncated approach." In: *Journal of Forecasting*.
In this study, we propose a new family of the heterogeneous autoregressive realized volatility (HAR-RV) models by considering truncated methods for predicting the RV in China's stock market. By adopting three types of critical values to recognize extremely large values of RV, we show that the modified models are simple but efficient to consistently deliver stronger in-sample and out-of-sample forecasting performances than those of existing methods. Models that take truncated approaches into account can generate substantial economic gains in applications. We further provide evidence that the superiority of our proposed models is derived from the reduced variance of the measurement errors during days including truncated RVs. Additionally, the improved performances of the modified models still hold after considering the effects of jump components and leverage, as well as a wide range of extensions and robustness analyses.

Westerlund, J., Karabiyik, H., and Narayan, P. (2017). "Testing for Predictability in panels with General Predictors." In: *Journal of Applied Econometrics* 32(3), pp. 554–574.
The difficulty of predicting returns has recently motivated researchers to start looking for tests that are either more powerful or robust to more features of the data. Unfortunately, the way that these tests work typically involves trading robustness for power or vice versa. The current paper takes this as its starting point to develop a new panel-based approach to predictability that is both robust and powerful. Specifically, while the panel route to increased power is not new, the way in which the cross-section variation is exploited also to achieve robustness with respect to the predictor is. The result is two new tests that enable asymptotically standard normal and chi-squared inference across a wide range of empirically relevant scenarios in which the predictor may be stationary, moderately non-stationary, nearly non-stationary, or indeed unit root non-stationary. The type of cross-section dependence that can be permitted in the predictor is also very general, and can be weak or strong, although we do require that the cross-section dependence in the regression errors is of the strong form. What is more, this generality comes at no cost in terms of complicated test construction. The new tests are therefore very user-friendly.

Winkler, R. L. (2015). "Equal Versus Differential Weighting in Combining Forecasts." In: *Risk Analysis* 35(11), pp. 16–18.

In conclusion, I will stick with my preference for simple combining rules such as a simple average and with the standard scoring rules developed in the literature. At the same time, I recognize that there are situations where differential weighting can be preferable. And having expressed my quibbles about the details of the calibration score and the weighting scheme in the classical method, I recognize that themethod is much more than just these elements, it has many positive features, and many users are evidently very satisfied.

Wu, H., Xu, J., Wang, J., and Long, M. (2022). "Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting." In: *arXiv e-Print*.

Extending the forecasting time is a critical demand for real applications, such as extreme weather early warning and long-term energy consumption planning. This paper studies the *long-term forecasting* problem of time series. Prior Transformer-based models adopt various self-attention mechanisms to discover the long-range dependencies. However, intricate temporal patterns of the long-term future prohibit the model from finding reliable dependencies. Also, Transformers have to adopt the sparse versions of point-wise self-attentions for long series efficiency, resulting in the information utilization bottleneck. Towards these challenges, we propose Autoformer as a novel decomposition architecture with an Auto-Correlation mechanism. We go beyond the pre-processing convention of series decomposition and renovate it as a basic inner block of deep models. This design empowers Autoformer with progressive decomposition capacities for complex time series. Further, inspired by the stochastic process theory, we design the Auto-Correlation mechanism based on the series periodicity, which conducts the dependencies discovery and representation aggregation at the sub-series level. Auto-Correlation outperforms self-attention in both efficiency and accuracy. In long-term forecasting, Autoformer yields state-of-the-art accuracy, with a 38% relative improvement on six benchmarks, covering five practical applications: energy, traffic, economics, weather and disease. Code is available at this repository: `https://github.com/thuml/Autoformer`.

Wu, Q., Brinton, C. G., Zhang, Z., Pizzoferrato, A., Liu, Z., and Cucuringu, M. (2021a). "Equity2Vec: End-to-end Deep Learning Framework for Cross-sectional Asset Pricing." In: *arXiv e-Print*.

Pricing assets has attracted significant attention from the financial technology community. We observe that the existing solutions overlook the cross-sectional effects and not fully leveraged the heterogeneous data sets, leading to sub-optimal performance. To this end, we propose an end-to-end deep learning framework to price the assets. Our framework possesses two main properties: 1) We propose Equity2Vec, a graph-based component that effectively captures both long-term and evolving cross-sectional interactions. 2) The framework simultaneously leverages all the available heterogeneous alpha sources including technical indicators, financial news signals, and cross-sectional signals. Experimental results on datasets from the real-world stock market show that our approach outperforms the existing state-of-the-art approaches. Furthermore, market trading simulations demonstrate that our framework monetizes the signals effectively.

Wu, X., Zhang, D., Guo, C., He, C., Yang, B., and Jensen, C. S. (2021b). "AutoCTS: Automated Correlated Time Series Forecasting – Extended Version." In: *arXiv e-Print*.

Correlated time series (CTS) forecasting plays an essential role in many cyber-physical systems, where multiple sensors emit time series that capture interconnected processes. Solutions based on deep learning that deliver state-of-the-art CTS forecasting performance employ a variety of spatio-temporal (ST) blocks that are able to model temporal dependencies and spatial correlations among time series. However, two challenges remain. First, ST-blocks are designed manually, which is time consuming and costly. Second, existing forecasting models simply stack the same ST-blocks multiple times, which limits the model potential. To address these challenges, we propose AutoCTS that is able to automatically identify highly competitive ST-blocks as well as forecasting models with heterogeneous ST-blocks connected using diverse topologies, as opposed to the same ST-blocks connected using simple stacking. Specifically, we design both a micro and a macro search space to model possible architectures of ST-blocks and the connections among heterogeneous ST-blocks, and we provide a search strategy that is able to jointly explore the search spaces to identify optimal forecasting models. Extensive experiments on eight commonly used CTS forecasting benchmark datasets justify our design choices and demonstrate that AutoCTS is capable of automatically discovering forecasting models that outperform state-of-the-art human-designed models. This is an extended version of "AutoCTS: Automated Correlated Time Series Forecasting", to appear in PVLDB 2022.

Xie, A. (2021). "Forecasting Long-Term Equity Returns: A Comparison of Popular Methodologies." In: *SSRN e-Print*.

Many investors need to make long-term asset class forecasts for planning and portfolio construction purposes. We examine the empirical performance of two different approaches to forecasting future ten-year equity returns: a regression methodology using CAPE and a more traditional ”building block” approach. The regression approach produces estimates that are poor predictors of subsequent actual returns. The "building block" approach (BBA) outperforms the regression methodology (in terms of root mean squared error) with the repricing component helping to capture periods of poor equity returns. A high CAPE value is not necessarily cause for alarm and changes in asset allocation. If an investor plans to use a methodology that over time will prove more accurate, then the historical record is more supportive of the BBA approach, with or without a repricing component based on current P/E.

Xu, W., Liu, W., Bian, J., Yin, J., and Liu, T.-Y. (2021). "Instance-wise Graph-based Framework for Multivariate Time Series Forecasting." In: *arXiv e-Print*.
The multivariate time series forecasting has attracted more and more attention because of its vital role in different fields in the real world, such as finance, traffic, and weather. In recent years, many research efforts have been proposed for forecasting multivariate time series. Although some previous work considers the interdependencies among different variables in the same timestamp, existing work overlooks the inter-connections between different variables at different time stamps. In this paper, we propose a simple yet efficient instance-wise graph-based framework to utilize the inter-dependencies of different variables at different time stamps for multivariate time series forecasting. The key idea of our framework is aggregating information from the historical time series of different variables to the current time series that we need to forecast. We conduct experiments on the Traffic, Electricity, and Exchange-Rate multivariate time series datasets. The results show that our proposed model outperforms the state-of-the-art baseline methods.

Xu, W., Liu, W., Wang, L., Xia, Y., Bian, J., Yin, J., and Liu, T.-Y. (2022). "HIST: A Graph-based Framework for Stock Trend Forecasting via Mining Concept-Oriented Shared Information." In: *arXiv e-Print*.
Stock trend forecasting, which forecasts stock prices' future trends, plays an essential role in investment. The stocks in a market can share information so that their stock prices are highly correlated. Several methods were recently proposed to mine the shared information through stock concepts (e.g., technology, Internet Retail) extracted from the Web to improve the forecasting results. However, previous work assumes the connections between stocks and concepts are stationary, and neglects the dynamic relevance between stocks and concepts, limiting the forecasting results. Moreover, existing methods overlook the invaluable shared information carried by hidden concepts, which measure stocks' commonness beyond the manually defined stock concepts. To overcome the shortcomings of previous work, we proposed a novel stock trend forecasting framework that can adequately mine the concept-oriented shared information from predefined concepts and hidden concepts. The proposed framework simultaneously utilize the stock's shared information and individual information to improve the stock trend forecasting performance. Experimental results on the real-world tasks demonstrate the efficiency of our framework on stock trend forecasting. The investment simulation shows that our framework can achieve a higher investment return than the baselines.

Yang, J.-Y., Zhu, H., Hou, Y.-J., Zhang, P., and Zhou, C.-C. (2021). "Why Existing Machine Learning Methods Fails At Extracting the Information of Future Returns Out of Historical Stock Prices : the Curve-Shape-Feature and Non-Curve-Shape-Feature Modes." In: *arXiv e-Print*.
The financial time series analysis is important access to touch the complex laws of financial markets. Among many goals of the financial time series analysis, one is to construct a model that can extract the information of the future return out of the known historical stock data, such as stock price, financial news, and e.t.c. To design such a model, prior knowledge on how the future return is correlated with the historical stock prices is needed. In this work, we focus on the issue: in what mode the future return is correlated with the historical stock prices. We manually design several financial time series where the future return is correlated with the historical stock prices in pre-designed modes, namely the curve-shape-feature (CSF) and the non-curve-shape-feature (NCSF) modes. In the CSF mode, the future return can be extracted from the curve shapes of the historical stock prices. By applying various kinds of existing algorithms on those pre-designed time series and real financial time series, we show that: (1) the major information of the future return is not contained in the curve-shape features of historical stock prices. That is, the future return is not mainly correlated with the historical stock prices in the CSF mode. (2) Various kinds of existing machine learning algorithms are good at extracting the curveshape features in the historical stock prices and thus are inappropriate for financial time series analysis although they are successful in the image recognition and natural language processing. That is, new models handling the NCSF series are needed in the financial time series analysis.

Yang, L., Li, J., Dong, R., Zhang, Y., and Smyth, B. (2022). "NumHTML: Numeric-Oriented Hierarchical Transformer Model for Multi-task Financial Forecasting." In: *arXiv e-Print*.

Financial forecasting has been an important and active area of machine learning research because of the challenges it presents and the potential rewards that even minor improvements in prediction accuracy or forecasting may entail. Traditionally, financial forecasting has heavily relied on quantitative indicators and metrics derived from structured financial statements. Earnings conference call data, including text and audio, is an important source of unstructured data that has been used for various prediction tasks using deep earning and related approaches. However, current deep learning-based methods are limited in the way that they deal with numeric data; numbers are typically treated as plain-text tokens without taking advantage of their underlying numeric structure. This paper describes a numeric-oriented hierarchical transformer model to predict stock returns, and financial risk using multi-modal aligned earnings calls data by taking advantage of the different categories of numbers (monetary, temporal, percentages etc.) and their magnitude. We present the results of a comprehensive evaluation of NumHTML against several state-of-the-art baselines using a real-world publicly available dataset. The results indicate that NumHTML significantly outperforms the current state-of-the-art across a variety of evaluation metrics and that it has the potential to offer significant financial gains in a practical trading context.

Yara, F. B., Boons, M., and Tamoni, A. (2021). "Value return predictability across asset classes and commonalities in risk premia." In: *Review of Finance* 25(2), pp. 449–484.

We show that returns to value strategies in individual equities, industries, commodities, currencies, global government bonds, and global stock indexes are predictable in the time series by their respective value spreads. In all these asset classes, expected value returns vary by at least as much as their unconditional level. A single common component of the value spreads captures about two-thirds of value return predictability and the remainder is asset class-specific. We argue that common variation in value premia is consistent with rationally time-varying expected returns, because (i) common value is closely associated with standard proxies for risk premia, such as the dividend yield, intermediary leverage, and illiquidity, and (ii) value premia are globally high in bad times.

Yeoleka, A., Patel, S., Talla, S., Puthucode, K. R., Ahmadzadeh, A., Sadykov, V. M., and Angryk, R. A. (2021). "Feature Selection on a Flare Forecasting Testbed: A Comparative Study of 24 Methods." In: *arXiv e-Print*.

The Space-Weather ANalytics for Solar Flares (SWAN-SF) is a multivariate time series benchmark dataset recently created to serve the heliophysics community as a testbed for solar flare forecasting models. SWAN-SF contains 54 unique features, with 24 quantitative features computed from the photospheric magnetic field maps of active regions, describing their precedent flare activity. In this study, for the first time, we systematically attacked the problem of quantifying the relevance of these features to the ambitious task of flare forecasting. We implemented an end-to-end pipeline for preprocessing, feature selection, and evaluation phases. We incorporated 24 Feature Subset Selection (FSS) algorithms, including multivariate and univariate, supervised and unsupervised, wrappers and filters. We methodologically compared the results of different FSS algorithms, both on the multivariate time series and vectorized formats, and tested their correlation and reliability, to the extent possible, by using the selected features for flare forecasting on unseen data, in univariate and multivariate fashions. We concluded our investigation with a report of the best FSS methods in terms of their top-k features, and the analysis of the findings. We wish the reproducibility of our study and the availability of the data allow the future attempts be comparable with our findings and themselves.

Yin, A. (2021). "Equity premium prediction: keep it sophisticatedly simple." In: *Quantitative Finance and Economics* 5(2), pp. 264–286.

Following the keep-it-sophisticatedly-simple principle, KISS, we propose using the averaging window approach to forecast the market equity premium in unstable environments. First, the estimation methodology of averaging window is a theoretically justified method robust to uncertainties on structural breaks and estimation window sizes. Second, the averaging window method has the obvious advantages of being understandable to forecast users and simple to implement, thus encouraging engagement and criticism. Our empirical results demonstrate the superior performance of the averaging window when forecasting the U.S. market equity premium, exceeding a wide range of methods which have been shown effective, such as shrinkage estimators and technical indicators.

Zang, C. (2017). "Deep Learning in Multiple Multistep Time Series Prediction." In: *arXiv e-Print*.

The project aims to research on combining deep learning specifically Long-Short Memory (LSTM) and basic statistics in multiple multistep time series prediction. LSTM can dive into all the pages and learn the general trends of variation in a large scope, while the well selected medians for each page can keep the special seasonality of different pages so that the future trend will not fluctuate too much from the reality. A recent Kaggle competition on 145K Web Traffic Time Series Forecasting [1] is used to thoroughly illustrate and test this idea.

Zeng, Z., Balch, T., and Veloso, M. (2021). "Deep Video Prediction for Time Series Forecasting." In: *arXiv e-Print*.

Time series forecasting is essential for decision making in many domains. In this work, we address the challenge of predicting prices evolution among multiple potentially interacting financial assets. A solution to this problem has obvious importance for governments, banks, and investors. Statistical methods such as Auto Regressive Integrated Moving Average (ARIMA) are widely applied to these problems. In this paper, we propose to approach economic time series forecasting of multiple financial assets in a novel way via video prediction. Given past prices of multiple potentially interacting financial assets, we aim to predict the prices evolution in the future. Instead of treating the snapshot of prices at each time point as a vector, we spatially layout these prices in 2D as an image, such that we can harness the power of CNNs in learning a latent representation for these financial assets. Thus, the history of these prices becomes a sequence of images, and our goal becomes predicting future images. We build on a state-of-the-art video prediction method for forecasting future images. Our experiments involve the prediction task of the price evolution of nine financial assets traded in U.S. stock markets. The proposed method outperforms baselines including ARIMA, Prophet, and variations of the proposed method, demonstrating the benefits of harnessing the power of CNNs in the problem of economic time series forecasting.

Zhan, T. and Xiao, F. (2021). "A Fast Evidential Approach for Stock Forecasting." In: *arXiv e-Print*.
In the framework of evidence theory, data fusion combines the confidence functions of multiple different information sources to obtain a combined confidence function. Stock price prediction is the focus of economics. Stock price forecasts can provide reference data. The Dempster combination rule is a classic method of fusing different information. By using the Dempster combination rule and confidence function based on the entire time series fused at each time point and future time points, and the preliminary forecast value obtained through the time relationship, the accurate forecast value can be restored. This article will introduce the prediction method of evidence theory. This method has good running performance, can make a rapid response on a large amount of stock price data, and has far-reaching significance.

Zhang, H. (2021). "Empirical asset pricing and ensemble machine learning." PhD thesis. Tilburg University.
Many of the sophisticated models for stock return forecasting and portfolio optimisation cannot beat naive equal-weighted models. The challenge is that, even in the age of big data, there are usually more potential variables than is appropriate for estimation. This thesis is dedicated to improving asset pricing models via ensemble machine learning methods without requiring more data. By introducing two ensemble methods, first, several representative sophisticated models of stock return forecasting are compared based on standard economic variables in the literature. The results show that both of the two ensemble methods could significantly improve these sophisticated models and found that these models can significantly outperform the equal-weighted combination of individual predictors. Their forecast gains stem from better performance during periods of market uncertainty and crises, and increased diversity and built-in shrinkage. Then, I introduce a general boosting framework for high-dimensional portfolio optimisation, where the classical mean-variance portfolios cannot work properly. The results indicate the effectiveness of these boosting methods in both low- and high-dimensional settings and they can outperform the 1/N portfolio in terms of several popular performance metrics.

Zhao, L. (2020). "Essays on Asset Pricing: A Model Comparison Perspective." PhD thesis. Washington University in St. Louis.
In my dissertation, I focus on theoretical and empirical asset pricing from a Bayesian model comparison perspective. In the first Chapter, revisiting the framework of Barillas and Shanken (2018), BS henceforth, we show that the Bayesian marginal likelihood-based model comparison method in that paper is unsound: the priors on the nuisance parameters across models must satisfy a change of variable property for densities that is violated by the Jeffreys priors used in the BS method. Extensive simulation exercises confirm that the BS method performs unsatisfactorily. We derive a new class of improper priors on the nuisance parameters, starting from a single improper prior, which leads to valid marginal likelihoods and model comparisons. The performance of our marginal likelihoods is significantly better, allowing for reliable Bayesian work on which factors are risk factors in asset pricing models. In the second Chapter, starting from the twelve distinct risk factors in four well-established asset pricing models, a pool we refer to as the winners, we construct and compare 4,095 asset pricing models and find that the model with the risk factors, Mkt, SMB, MOM, ROE, MGMT, and PEAD, performs the best in terms of Bayesian posterior probability, out-of-sample predictability, and Sharpe ratio. A more extensive model comparison of 8,388,607 models, constructed from the twelve winners plus eleven principal components of anomalies unexplained by the winners, shows the benefit of incorporating information in genuine anomalies in explaining the cross-section of expected equity returns.

Zhao, Y. (2021). "The robustness of forecast combination in unstable environments: a Monte Carlo study of advanced algorithms." In: *Empirical Economics* 61, pp. 173–199.

In this paper, we study the behavior and effectiveness of several recently developed forecast combination algorithms in simulated unstable environments, where the performances of individual forecasters are cross-sectionally heterogeneous and dynamically evolving. Our results clearly reveal how different algorithms respond to structural instabilities of different origin, frequency, and magnitude. Accordingly, we propose an improved forecast combination procedure and demonstrate its effectiveness in a real-time forecast combination exercise using the U.S. Survey of Professional Forecasters.

Zhao, Y., Wang, Y., Liu, J., Xia, H., Xu, Z., Hong, Q., Zhou, Z., and Petzold, L. (2021). "Empirical Quantitative Analysis of COVID-19 Forecasting Models." In: *arXiv e-Print*.

COVID-19 has been a public health emergency of international concern since early 2020. Reliable forecasting is critical to diminish the impact of this disease. To date, a large number of different forecasting models have been proposed, mainly including statistical models, compartmental models, and deep learning models. However, due to various uncertain factors across different regions such as economics and government policy, no forecasting model appears to be the best for all scenarios. In this paper, we perform quantitative analysis of COVID-19 forecasting of confirmed cases and deaths across different regions in the United States with different forecasting horizons, and evaluate the relative impacts of the following three dimensions on the predictive performance (improvement and variation) through different evaluation metrics: model selection, hyperparameter tuning, and the length of time series required for training. We find that if a dimension brings about higher performance gains, if not well-tuned, it may also lead to harsher performance penalties. Furthermore, model selection is the dominant factor in determining the predictive performance. It is responsible for both the largest improvement and the largest variation in performance in all prediction tasks across different regions. While practitioners may perform more complicated time series analysis in practice, they should be able to achieve reasonable results if they have adequate insight into key decisions like model selection.

Zhu, L., Basu, S., Jarrow, R. A., and Wells, M. T. (2021). "High-Dimensional Estimation, Basis Assets, and the Adaptive Multi-Factor Model." In: *arXiv e-Print*.

The paper proposes a new algorithm for the high-dimensional financial data – the Groupwise Interpretable Basis Selection (GIBS) algorithm, to estimate a new Adaptive Multi-Factor (AMF) asset pricing model, implied by the recently developed Generalized Arbitrage Pricing Theory, which relaxes the convention that the number of risk-factors is small. We first obtain an adaptive collection of basis assets and then simultaneously test which basis assets correspond to which securities, using high-dimensional methods. The AMF model, along with the GIBS algorithm, is shown to have a significantly better fitting and prediction power than the Fama-French 5-factor model.

Zhu, Y. and Timmermann, A. (2020). "Can Two Forecasts Have the Same Conditional Expected Accuracy?" In: *arXiv e-Print*.

The method for testing equal predictive accuracy for pairs of forecasting models proposed by Giacomini and White (2006) has found widespread use in empirical work. The procedure assumes that the parameters of the underlying forecasting models are estimated using a rolling window of fixed width and incorporates the effect of parameter estimation in the null hypothesis that two forecasts have identical conditionally expected loss. We show that this null hypothesis cannot be valid under a rolling window estimation scheme and even fails in the absence of parameter estimation for many types of stochastic processes in common use. This means that the approach does not guarantee appropriate comparisons of predictive accuracy of forecasting models. We also show that the Giacomini-White approach can lead to substantial size distortions in tests of equal unconditional predictive accuracy and propose an alternative procedure with better properties.

Ziel, F. and Berk, K. (2019). "Multivariate Forecasting Evaluation: On Sensitive and Strictly Proper Scoring Rules." In: *arXiv e-Print*.

In recent years, probabilistic forecasting is an emerging topic, which is why there is a growing need of suitable methods for the evaluation of multivariate predictions. We analyze the sensitivity of the most common scoring rules, especially regarding quality of the forecasted dependency structures. Additionally, we propose scoring rules based on the copula, which uniquely describes the dependency structure for every probability distribution with continuous marginal distributions. Efficient estimation of the considered scoring rules and evaluation methods such as the Diebold-Mariano test are discussed. In detailed simulation studies, we compare the performance of the renowned scoring rules and the ones we propose. Besides extended synthetic studies based on recently published results we also consider a real data example. We find that the energy score, which is probably the most widely used multivariate scoring rule, performs comparably well in detecting forecast errors, also regarding dependencies. This contradicts other studies. The results also show that a proposed copula score provides very strong distinction

between models with correct and incorrect dependency structure. We close with a comprehensive discussion on the proposed methodology.