

# References with abstracts for QWIM project: machine learning (and more) for scenario generation and data augmentation in quantitative wealth and investment management

Cristian Homescu

December 2022

## Abstract

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

## Contents

<b>1</b>	<b>Motivation for the project</b>	<b>2</b>
1.1	Need for additional data for QWIM	2
1.2	Using ML to generate synthetic data for QWIM	2
<b>2</b>	<b>Relevant references</b>	<b>4</b>
2.1	Main references	4
2.2	Comprehensive list of references	6
2.2.1	Scenarios incorporating interactions and dependencies across and within market periods	6
2.2.2	Scenarios based on subsets of factors	6
2.2.3	Incorporating expert views within scenarios	6
2.2.4	Data augmentation and scenarios generated using machine learning within context of QWIM	6
2.2.5	Scenarios based on flexible probabilities	7
2.2.6	Scenarios based on Monte Carlo simulation	7
2.2.7	Scenarios constructed using bootstrapping approach	8
2.2.8	Scenarios constructed using resampling approach	8
2.2.9	Scenarios based on similarities and discords within time series	8
2.2.10	Data augmentation and scenario generation using machine learning models	9
2.2.11	Scenarios for extreme events and stress testing	11
2.2.12	Scenarios for extreme events and stress testing within context of QWIM	11
2.2.13	Economic scenario generators	12
2.2.14	Scenarios incorporating dependencies (correlations, covariances, copulas)	13
2.2.15	Extreme events and scenarios	14
2.2.16	Extreme events and scenarios within context of QWIM	14
2.2.17	Scenarios based on agent-based modeling	14
2.2.18	Validating scenarios and augmented datasets	15
2.2.19	Software implementations and frameworks	15
	<b>References</b>	<b>17</b>

# 1 Motivation for the project

Statistical risk management approaches like Value at Risk (VaR) proved largely ineffective in addressing the kinds of market movements that occurred with past crisis periods. Scenario analysis addresses many of the shortcomings of the statistical approach, since it is both forward looking and also able to draw on historical experience that might not be represented in the development samples for statistical approaches.

There is a need to conduct rigorous and systematic scenario analysis. A key challenge in applying scenario analysis is constructing “tail event” scenarios that are severe but plausible. We can assess scenario’s plausibility by reference to the behavior of observable variables used to define the scenario (e.g., asset prices, bond yields, credit spreads, etc.). To generate realistic scenarios in all relevant market regimes we can use machine learning approaches to identify regimes.

Additional scenarios can also be generated through “data augmentation” algorithms used in machine learning (e.g., through transfer learning or generative adversarial networks GANs).

## 1.1 Need for additional data for QWIM

Not having sufficient data (due to medium to low sampling frequency and/or limited historical data) will have an impact on

- adequately training ML models
- tackling unbalanced datasets
- data anonymization and preserving of data privacy
- comprehensive testing of investment strategies and portfolios
- robust portfolio construction
- portfolio risk management

Synthetic data can be obtained through

- ML-based data augmentation **DAug**
- data similarity with data in other time periods
- combine observations with expert views
- agent based modeling
- scenario generation
- construction of stress scenarios

## 1.2 Using ML to generate synthetic data for QWIM

There is significant need for additional data in QWIM

- for adequately training ML models
- to tackle unbalanced datasets
- for constructing a very comprehensive set of scenarios incorporating most scenarios which were not observed, yet are plausible and consistent with market and investor behavior and with economic intuition
- for data anonymization and preserving of data privacy
- for comprehensive testing of investment strategies and portfolios
- for constructing much more robust portfolios using scenario-based optimization

- for better risk management of investment portfolios

Assefa et al. (“Generating Synthetic Data in Finance: Opportunities, Challenges and Pitfalls,” 2019):

synthetic data defined as data obtained from a generative process that learns properties of real data

Synthetic financial **TS** data is expected to have universal features, commonly referred to as stylized facts, together with specialized features of a specific asset class or investment vehicle. When classical approaches are considered, these stylized facts are often formulated in terms of distribution of returns.

Most promising **ML** methods to generate synthetic financial **TS**:

- Generative Adversarial Networks GANs
- Variational Autoencoders VAEs
- Restricted Boltzmann Machine RBM

## 2 Relevant references

### 2.1 Main references

List of references:

- Alaa et al. (“How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models,” 2022)
- Asch et al. (“Model-assisted deep learning of rare extreme events from partial observations,” 2022)
- Bahrpeyma et al. (“A Methodology for Validating Diversity in Synthetic Time Series Generation,” 2021)
- Bhatia et al. (“ExGAN: Adversarial Generation of Extreme Samples,” 2021)
- Bond-Taylor et al. (“Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models,” 2022)
- Bandara et al. (“Improving the accuracy of global forecasting models using time series data augmentation,” 2021)
- Buehler et al. (“A Data-driven Market Simulator for Small Data Environments,” 2020)
- Buehler et al. (“Generating financial markets with signatures,” 2021)
- Cerqueira et al. (“Evaluating time series forecasting models: an empirical study on performance estimation methods,” 2020)
- Cubuk et al. (“RandAugment: Practical data augmentation with no separate search,” 2019)
- Czasonis et al. (“Enhanced Scenario Analysis,” 2020)
- Czasonis et al. (“Relevance,” 2021)
- Dahl and Sorensen (“Time Series (re)sampling using Generative Adversarial Networks,” 2021)
- Da Silva and Shi (“Towards Improved Generalization in Financial Markets with Synthetic Data Generation,” 2019)
- Debnath et al. (“Exploring Generative Data Augmentation in Multivariate Time Series Forecasting : Opportunities and Challenges,” 2021)
- De Gennaro Aquino et al. (“Portfolio Selection With Exploration of New Investment Opportunities,” 2021)
- De Meer et al. (“Tackling the exponential scaling of signature-based GANs for high-dimensional financial time series generation,” 2021)
- De Meer Pardo (“Enriching Financial Datasets with Generative Adversarial Networks,” 2019)
- Diffenbaugh (“Verification of extreme event attribution: Using out-of-sample observations to assess changes in probabilities of unprecedented events.,” 2020)
- Ding et al. (“Modeling Extreme Events in Time Series Prediction,” 2019)
- Dogariu et al. (“Towards Realistic Financial Time Series Generation via Generative Adversarial Learning,” 2021)
- Dogariu et al. (“Generation of Realistic Synthetic Financial Time-Series,” 2021)
- Faggini et al. (“Crises in economic complex networks: Black Swans or Dragon Kings?” 2019)
- Flaig and Junike (“Scenario generation for market risk models using generative neural networks,” 2022)
- Franco-Pedroso et al. (“The ETS challenges: a machine learning approach to the evaluation of simulated financial time series for improving generation processes,” 2019)
- Franco-Pedroso et al. (“Generating Virtual Scenarios of Multivariate Financial Data for Quantitative Trading Applications,” 2019)
- Fons et al. (“Adaptive Weighting Scheme for Automatic Time-Series Data Augmentation,” 2021)
- Fons et al. (“Augmenting Transferred Representations for Stock Classification,” 2021)
- Fritzsche et al. (“Marginals Versus Copulas: Which Account For More Model Risk In Multivariate Risk Forecasting?” 2021)
- Fu et al. (“Data augmentation for time series,” 2020)
- Golub et al. (“Market-Driven Scenarios: An Approach for Plausible Scenario Construction,” 2018)
- He et al. (“Data Augmentation Revisited: Rethinking the Distribution Gap between Clean and Augmented Data,” 2019)
- Heaton and Witte (“Synthetic Financial Data: An Application to Regulatory Compliance for Broker-Dealers,” 2021)
- Kang et al. (“GRATIS: GeneRAtIng TIme Series with diverse and controllable characteristics,” 2020)
- Kang et al. (“Déjà vu: A data-centric forecasting approach through time series cross-similarity,” 2021)
- Kondratyev and Schwarz (“The Market Generator,” 2020)

Koshiyama et al. (“Generative Adversarial Networks for Financial Trading Strategies Fine-Tuning and Combination,” 2021)

Kritzman et al. (“Portfolio Choice with Path-Dependent Scenarios,” 2021)

Kritzman and Turkington (“History, Shocks, and Drifts: A New Approach to Portfolio Formation,” 2022)

Lerch et al. (“Forecaster’s Dilemma: Extreme Events and Forecast Evaluation,” 2017)

Lezmi et al. (“Improving the Robustness of Trading Strategy Backtesting with Boltzmann Machines and Generative Adversarial Networks,” 2020)

Lezmi et al. (“Portfolio Allocation with Skewness Risk: A Practical Guide,” 2019)

Li et al. (“SynC: A Copula based Framework for Generating Synthetic Data from Aggregated Sources,” 2020)

Liu et al. (“Efficient Time Series Augmentation Methods,” 2020)

Mannix and Cesa (“‘Signatures’ promise quants a tool for all jobs,” 2021)

Mariani et al. (“PAGAN: Portfolio Analysis with Generative Adversarial Networks,” 2019)

Marti et al. (“Exploring and measuring non-linear correlations: Copulas, Lightspeed Transportation and Clustering,” 2017)

Marti (“CORRGAN: sampling realistic financial correlation matrices using generative adversarial networks,” 2020)

Marti et al. (“cCorrGAN: Conditional Correlation GAN for Learning Empirical Conditional Distributions in the Elliptope,” 2021)

Kuchnik and Smith (“Efficient Augmentation via Data Subsampling,” 2018)

Packham and Woebbecking (“Correlation scenarios and correlation stress testing,” 2021)

Papenbrock et al. (“Matrix Evolutions: Synthetic Correlations and Explainable Machine Learning for Constructing Robust Investment Portfolios,” 2021)

Paul et al. (“PSA-GAN: Progressive Self Attention GANs for Synthetic Time Series,” 2022)

Pei et al. (“Towards Generating Real-World Time Series Data,” 2021)

Raab et al. (“Assessing, visualizing and improving the utility of synthetic data,” 2021)

Raghunathan (“Synthetic Data,” 2021)

Rosen and Saunders (“Regress under stress: A simple least-squares method for integrating economic scenarios with risk simulations,” 2016)

Rosen (“Integrating Economic Scenarios with Advanced Scenario Analytics to Manage Investment Portfolios,” 2015)

Rosen (“Re-Thinking Scenarios: Stress Testing of Multi-Asset Portfolios by Integrating Economic Scenarios with Advanced Simulation Analytics,” 2015)

Rosolia and Osterrieder (“Analyzing Deep Generated Financial Time Series for Various Asset Classes,” 2021)

Sahamkhadam and Stephan (“Portfolio Optimization Based on Forecasting Models Using Vine Copulas: An Empirical Assessment for the Financial Crisis,” 2021)

Sapp (“Efficient Estimation of Distributional Tail Shape and the Extremal Index with Applications to Risk Management,” 2017)

Sharma et al. (“V- U-, L- or W- shaped economic recovery after Covid-19: Insights from an Agent Based Model,” 2021)

Silva and Ferreira (“Surrogate Monte Carlo,” 2021)

Snow (“DataGene: A Framework for Dataset Similarity,” 2020)

Takahashi et al. (“Modeling financial time-series with generative adversarial networks,” 2019)

Wang et al. (“Scalable Data Augmentation for Deep Learning,” 2019)

Wen et al. (“Time Series Data Augmentation for Deep Learning: A Survey,” 2021)

Wiese et al. (“Quant GANs: deep generation of financial time series,” 2020)

Yuan and Yuan (“A Monte Carlo synthetic sample based performance evaluation method for covariance matrix estimators,” 2021)

Zhang et al. (“DADA: Deep Adversarial Data Augmentation for Extremely Low Data Regime Classification,” 2018)

Ziyin et al. (“What Data Augmentation Do We Need for Deep-Learning-Based Finance?” 2021)

Zorn (“Panic-aware portfolio optimization,” 2019)

## 2.2 Comprehensive list of references

### 2.2.1 Scenarios incorporating interactions and dependencies across and within market periods

List of references:

- Bass et al. (“Factor Performance Across Market-Driven Scenarios,” 2018)
- Czaronis et al. (“Enhanced Scenario Analysis,” 2020)
- Czaronis et al. (“Relevance,” 2021)
- Golub et al. (“Market-Driven Scenarios: An Approach for Plausible Scenario Construction,” 2018)
- Kritzman et al. (“Portfolio Choice with Path-Dependent Scenarios,” 2021)
- Kritzman and Turkington (“History, Shocks and Drifts: A New Approach to Portfolio Formation,” 2021)
- Papenbrock et al. (“Matrix Evolutions: Synthetic Correlations and Explainable Machine Learning for Constructing Robust Investment Portfolios,” 2021)

### 2.2.2 Scenarios based on subsets of factors

List of references:

- Glasserman et al. (“Stress scenario selection by empirical likelihood,” 2015)
- Rosen and Saunders (“Regress under stress: A simple least-squares method for integrating economic scenarios with risk simulations,” 2016)
- Rosen and Saunders (“Regress Under Stress: A Simple Least-Squares Method for Integrating Economic Scenarios with Risk Simulations,” 2015)
- Rosen (“Integrating Economic Scenarios with Advanced Scenario Analytics to Manage Investment Portfolios,” 2015)
- Rosen (“Re-Thinking Scenarios: Stress Testing of Multi-Asset Portfolios by Integrating Economic Scenarios with Advanced Simulation Analytics,” 2015)

### 2.2.3 Incorporating expert views within scenarios

List of references:

- Bolger and Wright (“Use of expert knowledge to anticipate the future: Issues, analysis and directions,” 2017)
- Colson and Cooke (“Cross validation for the classical model of structured expert judgment,” 2017)
- Davis and Lleo (“A Simple Procedure for Combining Expert Opinion with Statistical Estimates to Achieve Superior Portfolio Performance,” 2016)
- Davis and Lleo (“Behaviouralizing Black-Litterman: Expert Opinions and Behavioural Biases in a Diffusion Setting,” 2015)
- Franses and Bruijn (“Benchmarking Judgmentally Adjusted Forecasts,” 2017)
- Gzyl et al. (“Inferring probability densities from expert opinion,” 2017)
- Hsu et al. (“Bridging the divide in financial market forecasting: machine learners vs. financial economists,” 2016)
- Johnson and West (“Bayesian Predictive Synthesis: Forecast Calibration and Combination,” 2018)
- Nguyen and Chamroukhi (“Practical and theoretical aspects of mixture-of-experts modeling: An overview,” 2018)
- Pancheekha et al. (“Ensemble Active Management,” 2018)
- Silva et al. (“A more human-like portfolio optimization approach,” 2017)
- Wood (“How sure are we? Two approaches to statistical inference,” 2018)

### 2.2.4 Data augmentation and scenarios generated using machine learning within context of QWIM

List of references:

- Buehler et al. (“Deep hedging,” 2019)
- Buehler et al. (“A Data-driven Market Simulator for Small Data Environments,” 2020)
- Buehler et al. (“Generating financial markets with signatures,” 2021)
- Coletta et al. (“Towards Realistic Market Simulations: a Generative Adversarial Networks Approach,” 2021)
- De Meer Pardo (“Enriching Financial Datasets with Generative Adversarial Networks,” 2019)

De Meer Pardo and Lopez (“Mitigating Overfitting on Financial Datasets with Generative Adversarial Networks,” 2020)

de Miranda Cardoso et al. (“Algorithms for Learning Graphs in Financial Markets,” 2020)

Eckerli (“Generative Adversarial Networks in finance: an overview,” 2021)

Fons et al. (“Adaptive Weighting Scheme for Automatic Time-Series Data Augmentation,” 2021)

Fons et al. (“Augmenting Transferred Representations for Stock Classification,” 2021)

Franco-Pedroso et al. (“Generating Virtual Scenarios of Multivariate Financial Data for Quantitative Trading Applications,” 2019)

Franco-Pedroso et al. (“The ETS challenges: a machine learning approach to the evaluation of simulated financial time series for improving generation processes,” 2019)

Heaton and Witte (“Synthetic Financial Data: An Application to Regulatory Compliance for Broker-Dealers,” 2021)

Henry-Labordere (“Generative models for financial data,” 2019)

Kondratyev and Schwarz (“The Market Generator,” 2020)

Koshiyama et al. (“Generative Adversarial Networks for Financial Trading Strategies Fine-Tuning and Combination,” 2021)

Lezmi et al. (“Improving the Robustness of Trading Strategy Backtesting with Boltzmann Machines and Generative Adversarial Networks,” 2020)

Marti (“CORRGAN: sampling realistic financial correlation matrices using generative adversarial networks,” 2020)

Marti (“Generating Realistic Synthetic Data in Finance: Applications of GANs,” 2020)

Rikli et al. (“Wasserstein GAN: Deep Generation applied on Bitcoins financial time series,” 2021)

Rosolia and Osterrieder (“Analyzing Deep Generated Financial Time Series for Various Asset Classes,” 2021)

Shah and Shroff (“Forecasting Market Prices using DL with Data Augmentation and Meta-learning: ARIMA still wins!” 2021)

Takahashi et al. (“Modeling financial time-series with generative adversarial networks,” 2019)

Wiese et al. (“Quant GANs: deep generation of financial time series,” 2020)

Ziyin et al. (“What Data Augmentation Do We Need for Deep-Learning-Based Finance?” 2021)

### 2.2.5 Scenarios based on flexible probabilities

List of references:

Ardia and Meucci (“Stress testing in non-normal markets via entropy pooling,” 2015)

Ardia and Bluteau (“Stress-Testing With Parametric Models and Fully Flexible Probabilities,” 2017)

Meucci (“Fully Flexible Views: Theory and Practice,” 2008)

Meucci et al. (“Fully Flexible Extreme Views,” 2012)

Sebastian and Gebbie (“Systematic Asset Allocation using Flexible Views for South African Markets,” 2019)

### 2.2.6 Scenarios based on Monte Carlo simulation

List of references:

Chen and Chen (*Monte-Carlo Simulation-Based Statistical Modeling*, 2017)

Juan et al. (“A review of the role of heuristics in stochastic optimisation: from metaheuristics to learnheuristics,” 2022)

Lam and Li (“Parametric Scenario Optimization under Limited Data: A Distributionally Robust Optimization View,” 2020)

Ledermann et al. (“Random orthogonal matrix simulation,” 2011)

Ledermann and Alexander (“Further properties of random orthogonal matrix simulation,” 2012)

Lee (“Stress testing Monte Carlo assumptions,” 2013)

Leon and Reveiz (“Monte Carlo Simulation of Long-Term Dependent Processes: A Primer,” 2012)

Schissler et al. (“Simulating High-Dimensional Multivariate Data using the bigsimr R Package,” 2021)

Silva and Ferreira (“Surrogate Monte Carlo,” 2021)

Staum (“Monte Carlo Computation in Finance,” 2009)

Yuan and Yuan (“A Monte Carlo synthetic sample based performance evaluation method for covariance matrix estimators,” 2021)



Zhang (“Modern Monte Carlo methods for efficient uncertainty quantification and propagation: A survey,” 2020)

### 2.2.7 Scenarios constructed using bootstrapping approach

List of references:

- Bertsimas and Van Parys (“Bootstrap Robust Prescriptive Analytics,” 2017)
- Cavaliere et al. (“An Introduction to Bootstrap Theory in Time Series Econometrics,” 2020)
- Cogneau and Zakamouline (“Block bootstrap methods and the choice of stocks for the long run,” 2013)
- Davidson (“Diagnostics for the bootstrap and fast double bootstrap,” 2017)
- El Karoui and Purdom (“Can we trust the bootstrap in high-dimensions? the case of linear models,” 2018)
- Gilleland (“Bootstrap Methods for Statistical Inference. Part I: Comparative Forecast Verification for Continuous Variables,” 2020)
- Gilleland (“Bootstrap Methods for Statistical Inference. Part II: Extreme-Value Analysis,” 2020)
- Goncalves and Perron (“Bootstrapping factor models with cross sectional dependence,” 2020)
- Hambuckers and Heuchenne (“Estimating the Out-of-Sample Predictive Ability of Trading Rules: A Robust Bootstrap Approach,” 2016)
- Honore and Hu (“Poor (Wo)man’s Bootstrap,” 2017)
- Horowitz (“Bootstrap methods in econometrics,” 2019)
- Imbens and Menzel (“A causal bootstrap,” 2021)
- Jaeger et al. (“Understanding machine learning for diversified portfolio construction by explainable AI,” 2020)
- Kreiss and Paparoditis (“Bootstrap methods for dependent data: A review,” 2011)
- Little and Badawy (“Causal bootstrapping,” 2020)
- Page (“How to combine long and short return histories efficiently,” 2013)
- Pathak and Rao (“The Sequential Bootstrap,” 2013)
- Romano and Wolf (“Multiple Testing of One-Sided Hypotheses: Combining Bonferroni and the Bootstrap,” 2018)
- Sani et al. (“The replacement bootstrap for dependent data,” 2015)
- Tsamardinos et al. (“Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation,” 2018)
- Vinod (“Constructing Scenarios of Time Heterogeneous Series for Stress Testing,” 2012)
- Wang and Tu (“Bootstrap methods: the classical theory and recent development,” 2014)
- Xu et al. (“Applications of the Fractional-Random-Weight Bootstrap,” 2020)

### 2.2.8 Scenarios constructed using resampling approach

List of references:

- Cerqueira et al. (“Evaluating time series forecasting models: an empirical study on performance estimation methods,” 2020)
- Chihara and Hesterberg (*Mathematical Statistics with Resampling and R, 2nd Edition*, 2018)
- Dahl and Sorensen (“Time Series (re)sampling using Generative Adversarial Networks,” 2021)
- Frahm (“A theoretical foundation of portfolio resampling,” 2015)
- Hoffman (“Resampling Statistics,” 2015)
- Huang and Yu (“A new procedure for resampled portfolio with shrinkaged covariance matrix,” 2020)
- Mammen and Nandi (“Bootstrap and Resampling,” 2012)
- Michaud and Michaud (“Estimation Error and Portfolio Optimization: A Resampling Solution,” 2015)
- Oneto (“Resampling Methods,” 2019)
- Sarris et al. (“Exploiting resampling techniques for model selection in forecasting: an empirical evaluation using out-of-sample tests,” 2020)

### 2.2.9 Scenarios based on similarities and discords within time series

List of references:

- Gharghabi et al. (“Matrix profile XII: mpdist: A novel time series distance measure to allow data mining in more challenging scenarios,” 2018)
- Gonzalez et al. (“Similarity Metrics for Transfer Learning in Financial Markets,” 2021)



- Kang et al. (“Déjà vu: A data-centric forecasting approach through time series cross-similarity,” 2021)
- Kegel et al. (“Feature-based comparison and generation of time series,” 2018)
- Kegel et al. (“Generating What-If Scenarios for Time Series Data,” 2017)
- Linardi et al. (“VALMOD: A suite for easy and exact detection of variable length motifs in data series,” 2018)
- Mikalsen et al. (“Time Series Cluster Kernel for Learning Similarities between Multivariate Time Series with Missing Data,” 2018)
- Snow (“DataGene: A Framework for Dataset Similarity,” 2020)
- Stephanovitch et al. (“Optimal 1-Wasserstein Distance for WGANs,” 2022)
- Yeh et al. (“Time series joins, motifs, discords and shapelets: a unifying view that exploits the matrix profile,” 2017)
- Yeh (“Towards a Near Universal Time Series Data Mining Tool: Introducing the Matrix Profile,” 2020)
- Zhu et al. (“Time series chains: A novel tool for time series data mining,” 2018)
- Zhu et al. (“Introducing time series chains: a new primitive for time series data mining,” 2019)

### 2.2.10 Data augmentation and scenario generation using machine learning models

List of references:

- Bandara et al. (“Improving the accuracy of global forecasting models using time series data augmentation,” 2021)
- Bhattacharai et al. (“Sampling Strategies for GAN Synthetic Data,” 2020)
- Bond-Taylor et al. (“Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models,” 2022)
- Chadebec and Allasonniere (“Data Augmentation with Variational Autoencoders and Manifold Sampling,” 2021)
- Chalongvorachai and Woraratpanya (“A data generation framework for extremely rare case signals,” 2021)
- Charte et al. (“A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines,” 2018)
- Chen et al. (“An Empirical Survey of Data Augmentation for Limited Data Learning in NLP,” 2021)
- Cao and Guo (“Generative Adversarial Network: Some Analytical Perspectives,” 2021)
- Dahl and Sorensen (“Time Series (re)sampling using Generative Adversarial Networks,” 2021)
- Debnath et al. (“Exploring Generative Data Augmentation in Multivariate Time Series Forecasting : Opportunities and Challenges,” 2021)
- De Meer et al. (“Tackling the exponential scaling of signature-based GANs for high-dimensional financial time series generation,” 2021)
- Dogariu et al. (“Towards Realistic Financial Time Series Generation via Generative Adversarial Learning,” 2021)
- Dogariu et al. (“Generation of Realistic Synthetic Financial Time-Series,” 2021)
- Eckerli and Osterrieder (“Generative Adversarial Networks in finance: an overview,” 2021)
- Faez et al. (“Deep Graph Generators: A Survey,” 2020)
- Fakoor et al. (“Fast, Accurate, and Simple Models for Tabular Data via Augmented Distillation,” 2020)
- Fang and Lin (“Prior knowledge distillation based on financial time series,” 2020)
- Fawaz et al. (“Data augmentation using synthetic data for time series classification with deep residual networks,” 2018)
- Fons et al. (“Adaptive Weighting Scheme for Automatic Time-Series Data Augmentation,” 2021)
- Fu et al. (“Data augmentation for time series,” 2020)
- Goel et al. (“Model Patching: Closing the Subgroup Performance Gap with Data Augmentation,” 2020)
- He et al. (“Data Augmentation Revisited: Rethinking the Distribution Gap between Clean and Augmented Data,” 2019)
- Hofert et al. (“Multivariate time-series modeling with generative neural networks,” 2021)
- Hoffmann et al. (“Machine Learning in a data-limited regime: Augmenting experiments with synthetic data uncovers order in crumpled sheets,” 2018)
- Imokoyende (“Variational Autoencoder In Finance,” 2019)
- Iwana and Uchida (“An empirical survey of data augmentation for time series classification with neural networks,” 2021)
- Jabbar et al. (“A Survey on Generative Adversarial Networks: Variants, Applications, and Training,” 2020)

Javeri et al. (“Improving Neural Networks for Time-Series Forecasting using Data Augmentation and AutoML,” 2021)

Jiang et al. (“Deceive D: Adaptive Pseudo Augmentation for GAN Training with Limited Data,” 2021)

Jin and Rinard (“Towards Context-Agnostic Learning Using Synthetic Data,” 2021)

Kingma and Welling (“An introduction to variational autoencoders,” 2019)

Koesdwiady et al. (“Methods to Improve Multi-Step Time Series Prediction,” 2018)

Kong et al. (“Robust Optimization as Data Augmentation for Large-scale Graphs,” 2022)

Koochali et al. (“If You Like It, GAN It. Probabilistic Multivariate Times Series Forecast With GAN,” 2020)

Kuchnik and Smith (“Efficient Augmentation via Data Subsampling,” 2018)

Laskin et al. (“Reinforcement Learning with Augmented Data,” 2020)

Lawrence et al. (“Data Generating Process to Evaluate Causal Discovery Techniques for Time Series Data,” 2021)

Lezmi et al. (“Improving the Robustness of Trading Strategy Backtesting with Boltzmann Machines and Generative Adversarial Networks,” 2020)

Leznik et al. (“Multivariate Time Series Synthesis Using Generative Adversarial Networks,” 2021)

Li et al. (“Improving GAN with inverse cumulative distribution function for tabular data synthesis,” 2021)

Li et al. (“A Synthetic Regression Model for Large Portfolio Allocation,” 2022)

Lim et al. (“Noisy Feature Mixup,” 2021)

Liu et al. (“Efficient Time Series Augmentation Methods,” 2020)

Mannix and Cesa (“‘Signatures’ promise quants a tool for all jobs,” 2021)

Ni et al. (“Conditional Sig-Wasserstein GANs for Time Series Generation,” 2020)

Padhi et al. (“Tabular Transformers for Modeling Multivariate Time Series,” 2021)

Pan et al. (“Data-Centric Engineering: integrating simulation, machine learning and statistics. Challenges and Opportunities,” 2021)

Pei et al. (“Towards Generating Real-World Time Series Data,” 2021)

Perez et al. (“Data augmentation through multivariate scenario forecasting in Data Centers using Generative Adversarial Networks,” 2022)

Raileanu et al. (“Automatic Data Augmentation for Generalization in Deep Reinforcement Learning,” 2021)

Rajabi and Garibay (“TabFairGAN: Fair Tabular Data Generation with Generative Adversarial Networks,” 2021)

Rocca and Rocca (“Intuitively Understanding Variational Autoencoders,” 2019)

Rosolia and Osterrieder (“Analyzing Deep Generated Financial Time Series for Various Asset Classes,” 2021)

Salazar et al. (“Generative Adversarial Networks and Markov Random Fields for oversampling very small training sets,” 2021)

Saxena and Cao (“Generative Adversarial Networks (GANs),” 2022)

Saxena and Cao (“Generative Adversarial Networks (GANs): Challenges, Solutions, and Future Directions,” 2020)

Shah and Shroff (“Forecasting Market Prices using DL with Data Augmentation and Meta-learning: ARIMA still wins!” 2021)

Shen et al. (“Towards Out-Of-Distribution Generalization: A Survey,” 2021)

Smith and Smith (“Conditional GAN for timeseries generation,” 2020)

Sun et al. (“Decision-Aware Conditional GANs for Time Series Data,” 2020)

Taylor and Nitschke (“Improving Deep Learning using Generic Data Augmentation,” 2017)

Tran et al. (“A Bayesian Data Augmentation Approach for Learning Deep Models,” 2017)

Vieira (“Generating Synthetic Sequential Data using GANs,” 2020)

Volpi et al. (“Generalizing to Unseen Domains via Adversarial Data Augmentation,” 2018)

Wang et al. (“Regularizing Deep Networks with Semantic Data Augmentation,” 2020)

Wang et al. (“Scalable Data Augmentation for Deep Learning,” 2019)

Wen et al. (“Time Series Data Augmentation for Deep Learning: A Survey,” 2021)

Xie et al. (“Unsupervised Data Augmentation,” 2019)

Yacoby et al. (“Characterizing and Avoiding Problematic Global Optima of Variational Autoencoders,” 2020)

Yao et al. (“Learning with Small Data,” 2020)

Ye and Dai (“Implementing transfer learning across different datasets for time series forecasting,” 2021)

Yoon et al. (“Time-series Generative Adversarial Networks,” 2019)

Xu et al. (“Unsupervised meta-learning for few-shot learning,” 2021)  
 Yu (“A Tutorial on VAEs: From Bayes’ Rule to Lossless Compression,” 2020)  
 Zhu et al. (“Pagan: Portfolio Analysis with Generative Adversarial Networks,” 2021)

### 2.2.11 Scenarios for extreme events and stress testing

List of references:

Bhatia et al. (“ExGAN: Adversarial Generation of Extreme Samples,” 2021)  
 Chalongvorachai and Woraratpanya (“A data generation framework for extremely rare case signals,” 2021)  
 Diffenbaugh (“Verification of extreme event attribution: Using out-of-sample observations to assess changes in probabilities of unprecedented events.,” 2020)  
 Ding et al. (“Modeling Extreme Events in Time Series Prediction,” 2019)  
 Engelke and Ivanovs (“Sparse Structures for Multivariate Extremes,” 2021)  
 Lerch et al. (“Forecaster’s Dilemma: Extreme Events and Forecast Evaluation,” 2017)  
 Naveau et al. (“Statistical methods for extreme event attribution in climate science,” 2020)  
 Qi and Majda (“Using machine learning to predict extreme events in complex systems.,” 2019)  
 Schwaab et al. (“Modeling Extreme Events: Time-Varying Extreme Tail Shape,” 2021)  
 Zhao (“Event Prediction in the Big Data Era,” 2022)

### 2.2.12 Scenarios for extreme events and stress testing within context of QWIM

List of references:

Al Janabi et al. (“Multivariate dependence and portfolio optimization algorithms under illiquid market scenarios,” 2017)  
 Albanese et al. (“Reverse Stress Testing,” 2020)  
 Alokley and Albarrak (“Clustering of Extremes in Financial Returns: A Study of Developed and Emerging Markets,” 2020)  
 Ardia and Meucci (“Stress testing in non-normal markets via entropy pooling,” 2015)  
 Ardia and Bluteau (“Stress-Testing With Parametric Models and Fully Flexible Probabilities,” 2017)  
 Arora et al. (“Extreme Yet Plausible: Choosing Scenarios to Stress Test Financial Institutions,” 2021)  
 Aste (“Stress testing and systemic risk measures using multivariate conditional probability,” 2021)  
 Baes and Schaanning (“Reverse Stress Testing: Scenario Design for Macroprudential Stress Tests,” 2020)  
 Breuer and Summer (“Systematic stress tests on public data,” 2020)  
 Bianchi et al. (*Handbook of Heavy-Tailed Distributions in Asset Management and Risk Management*, 2019)  
 Bilgili et al. (“Stress Hedging in Portfolio Construction,” 2017)  
 Boubaker and Nguyen (*Handbook of Global Financial Markets*, 2019)  
 Chalkis et al. (“Modeling of crisis periods in stock markets,” 2021)  
 Chen and Nasekin (“Quantifying systemic risk with factor copulas,” 2020)  
 Cohort et al. (“Analytical scores for stress scenarios,” 2020)  
 Denev and Mutnikas (“A Formalized, Integrated and Visual Approach to Stress Testing,” 2016)  
 Engle (“Stress Testing with Market Data,” 2020)  
 Flood and Korenko (“Systematic scenario selection: stress testing and the nature of uncertainty,” 2015)  
 Gao et al. (“Efficient Simulation of Financial Stress Testing Scenarios with Suppes-Bayes Causal Networks,” 2017)  
 Gao et al. (“Causal data science for financial stress testing,” 2018)  
 Glasserman et al. (“Stress scenario selection by empirical likelihood,” 2015)  
 Grundke (“Further recipes for quantitative reverse stress testing,” 2012)  
 Kantos and diBartolomeo (“How the pandemic taught us to turn smart beta into real alpha,” 2020)  
 Kopeliovich et al. (“Robust Risk Estimation and Hedging: A Reverse Stress Testing Approach,” 2015)  
 Meucci (*A Fully Integrated Liquidity and Market Risk Model*, 2012)  
 Pesenti et al. (“Scenario weights for importance measurement (SWIM) - an R package for sensitivity analysis,” 2020)  
 Pesenti et al. (“Reverse sensitivity testing: What does it take to break the model?” 2019)  
 Rebonato (“The quickest way to lose the money you cannot afford to lose: reverse stress testing with maximum entropy,” 2018)

Rojas and Dias (“Stress testing network reconstruction via graphical causal model,” 2021)  
 Rosen and Saunders (“Regress under stress: A simple least-squares method for integrating economic scenarios with risk simulations,” 2016)  
 Ruenzi et al. (“Joint Extreme events in equity returns and liquidity and their cross-sectional pricing implications,” 2020)  
 Schwaab et al. (“Modeling Extreme Events: Time-Varying Extreme Tail Shape,” 2021)  
 Seabrook et al. (“An Information Filtering approach to stress testing: an application to FTSE markets,” 2021)  
 Siddique et al. (*Stress Testing (Second Edition)*, 2019)  
 Skoglund (“Quantification of model risk in stress testing and scenario analysis,” 2019)  
 Tanaka (“Forecasting scenarios from the perspective of a reverse stress test using second-order cone programming,” 2017)  
 Traccucci et al. (“A Triptych Approach for Reverse Stress Testing of Complex Portfolios,” 2019)  
 Wang et al. (“Forecasting stock volatility in the presence of extreme shocks: Short-term and long-term effects,” 2020)  
 Zorn (“Panic-aware portfolio optimization,” 2019)

### 2.2.13 Economic scenario generators

List of references:

Barra Montevechi et al. (“A simulation-based approach to perform economic evaluation scenarios,” 2017)  
 Cheng and Planchet (“Stochastic Deflator for an Economic Scenario Generator with Five Factors,” 2018)  
 Chen et al. (“Personalised drawdown strategies and partial annuitisation to mitigate longevity risk,” 2021)  
 Chen et al. (“Using a stochastic economic scenario generator to analyse uncertain superannuation and retirement outcomes,” 2021)  
 Czaronis et al. (“Enhanced Scenario Analysis,” 2020)  
 De Meo (“Scenario Design for Macro-Financial Stress Testing,” 2019)  
 Facchinato and Pola (“Managing uncertainty with diversification across macroeconomic scenarios (DAMS): from asset segmentation to portfolio,” 2014)  
 Flaig and Junike (“Scenario generation for market risk models using generative neural networks,” 2022)  
 Golub et al. (“Market-Driven Scenarios: An Approach for Plausible Scenario Construction,” 2018)  
 Haldane and Turrell (“Drawing on different disciplines: macroeconomic agent-based models,” 2018)  
 Lee et al. (“How can an economic scenario generation model cope with abrupt changes in financial markets?” 2021)  
 Moudiki and Planchet (“Economic Scenario Generators,” 2016)  
 Ortec Finance (“Ex ante risk management with scenarios,” 2014)  
 Rosen (“Integrating Economic Scenarios with Advanced Scenario Analytics to Manage Investment Portfolios,” 2015)  
 Rosen (“Re-Thinking Scenarios: Stress Testing of Multi-Asset Portfolios by Integrating Economic Scenarios with Advanced Simulation Analytics,” 2015)  
 Rosen and Saunders (“Regress Under Stress: A Simple Least-Squares Method for Integrating Economic Scenarios with Risk Simulations,” 2015)  
 Rosen and Saunders (“Regress under stress: A simple least-squares method for integrating economic scenarios with risk simulations,” 2016)  
 Schneider (“Sparse economic scenarios,” 2020)  
 Sharma et al. (“V- U-, L- or W- shaped economic recovery after Covid-19: Insights from an Agent Based Model,” 2021)  
 Siew et al. (“The Impact of Different Economic Scenarios Towards Portfolio Selection in Enhanced Index Tracking Problem,” 2015)  
 Steehouwer (“Ex-ante risk management with scenarios,” 2014)  
 Steehouwer and Slater (“Macroeconomic Scenarios: A Frequency Domain Approach,” 2010)  
 Trimborn et al. (“SABCEMM: A Simulator for Agent-Based Computational Economic Market Models,” 2020)  
 van Beek (“Consistent Calibration of Economic Scenario Generators: The Case for Conditional Simulation,” 2020)  
 Wang (“Discriminating modelling approaches for Point in Time Economic Scenario Generation,” 2021)

Zhang (“Optimal Retirement Planning: Scenario Generation, Preferences, and Objectives,” 2018)

#### 2.2.14 Scenarios incorporating dependencies (correlations, covariances, copulas)

List of references:

- Alexander and Ledermann (“ROM Simulation: Applications to Stress Testing and VaR,” 2012)
- Amengual and Sentana (“Is a Normal Copula the Right Copula?” 2020)
- Benali et al. (“MTCopula: Synthetic Complex Data Generation Using Copula,” 2021)
- Carrillo et al. (“A New Machine Learning Forecasting Algorithm Based on Bivariate Copula Functions,” 2021)
- Czado and Nagler (“Vine Copula Based Modeling,” 2022)
- Fritzsch et al. (“Marginals Versus Copulas: Which Account For More Model Risk In Multivariate Risk Forecasting?” 2021)
- Fulle and Herwartz (“A Multivariate Markov-Switching GARCH Model with Copula-Distributed Innovations,” 2021)
- Galeeva et al. (“Measuring correlation risk for Energy Derivatives,” 2012)
- Großer and Okhrin (“Copulae: An overview and recent developments,” 2021)
- Harutyunyan et al. (“Efficient Covariance Estimation from Temporal Data,” 2021)
- Janke et al. (“Implicit Generative Copulas,” 2021)
- Ledermann and Alexander (“Further properties of random orthogonal matrix simulation,” 2012)
- Ledoit and Wolf (“Shrinkage estimation of large covariance matrices: Keep it simple, statistician?” 2021)
- Li et al. (“SynC: A Copula based Framework for Generating Synthetic Data from Aggregated Sources,” 2020)
- Lu and Ghosh (“Nonparametric estimation of multivariate copula using empirical bayes method,” 2021)
- Ma (“copent: Estimating Copula Entropy and Transfer Entropy in R,” 2021)
- Marti et al. (“Exploring and measuring non-linear correlations: Copulas, Lightspeed Transportation and Clustering,” 2017)
- Marti (“CORRGAN: sampling realistic financial correlation matrices using generative adversarial networks,” 2020)
- Marti et al. (“cCorrGAN: Conditional Correlation GAN for Learning Empirical Conditional Distributions in the Elliptope,” 2021)
- Marti et al. (“A review of two decades of correlations, hierarchies, networks and clustering in financial markets,” 2021)
- Meyer and Nagler (“Synthia: multidimensional synthetic data generation in Python,” 2021)
- Nasri and Rémillard (“Copula-based dynamic models for multivariate time series,” 2019)
- Nasri et al. (“Goodness-of-fit for regime-switching copula models with application to option pricing,” 2020)
- Opdyke (“Full Probabilistic Control for Direct and Robust, Generalized and Targeted Stressing of the Correlation Matrix (Even When Eigenvalues are Empirically Challenging,” 2020)
- Packham and Woebeking (“Correlation scenarios and correlation stress testing,” 2021)
- Papenbrock et al. (“Matrix Evolutions: Synthetic Correlations and Explainable Machine Learning for Constructing Robust Investment Portfolios,” 2021)
- Sahamkhadam and Stephan (“Portfolio Optimization Based on Forecasting Models Using Vine Copulas: An Empirical Assessment for the Financial Crisis,” 2021)
- Tan and Zohren (“Large Non-Stationary Noisy Covariance Matrices: A Cross-Validation Approach,” 2021)
- Traccucci et al. (“A Triptych Approach for Reverse Stress Testing of Complex Portfolios,” 2019)
- Trucíos et al. (“Forecasting Conditional Covariance Matrices in High-Dimensional Time Series: a General Dynamic Factor Approach,” 2021)
- Tsanakas and Zhu (“Copula model selection using image recognition,” 2021)
- Waller (“Generating Correlation Matrices With Specified Eigenvalues Using the Method of Alternating Projections,” 2020)
- Yu et al. (“Adjusting covariance matrix for risk management,” 2020)
- Yuan and Yuan (“A Monte Carlo synthetic sample based performance evaluation method for covariance matrix estimators,” 2021)



### 2.2.15 Extreme events and scenarios

#### References:

- Asch et al. (“Model-assisted deep learning of rare extreme events from partial observations,” 2022)
- Bhatia et al. (“ExGAN: Adversarial Generation of Extreme Samples,” 2021)
- Buczak et al. (“Crystal Cube: Forecasting Disruptive Events,” 2022)
- Chevalier et al. (“Modeling Nonstationary Extreme Dependence With Stationary Max-Stable Processes and Multidimensional Scaling,” 2021)
- Diffenbaugh (“Verification of extreme event attribution: Using out-of-sample observations to assess changes in probabilities of unprecedented events.,” 2020)
- Ding et al. (“Modeling Extreme Events in Time Series Prediction,” 2019)
- Faggini et al. (“Crises in economic complex networks: Black Swans or Dragon Kings?” 2019)
- Lerch et al. (“Forecaster’s Dilemma: Extreme Events and Forecast Evaluation,” 2017)
- Medina et al. (“Spectral learning of multivariate extremes,” 2021)
- Naveau et al. (“Statistical methods for extreme event attribution in climate science,” 2020)
- Qi and Majda (“Using machine learning to predict extreme events in complex systems.,” 2019)
- Racca and Magri (“Statistical prediction of extreme events from small datasets,” 2022)
- Yilmaz et al. (“Comparison of different estimation methods for extreme value distribution,” 2021)

### 2.2.16 Extreme events and scenarios within context of QWIM

#### References:

- Cai et al. (“New volatility evolution model after extreme events,” 2022)
- De Luca and Zuccolotto (“A double clustering algorithm for financial time series based on extreme events,” 2016)
- Elkamhi and Stefanova (“Dynamic Hedging and Extreme Asset Co-movements,” 2015)
- Kapoor and Shrivastava (“Extreme Values Theory and Return Level Analysis for Catastrophe Prediction,” 2014)
- Kemp (*Extreme Events: Robust Portfolio Construction in the Presence of Fat Tails*, 2014)
- Lezmi et al. (“Portfolio Allocation with Skewness Risk: A Practical Guide,” 2019)
- Ruenzi et al. (“Joint Extreme events in equity returns and liquidity and their cross-sectional pricing implications,” 2020)
- Sapp (“Efficient Estimation of Distributional Tail Shape and the Extremal Index with Applications to Risk Management,” 2017)
- Zorn (“Panic-aware portfolio optimization,” 2019)

### 2.2.17 Scenarios based on agent-based modelling

#### List of references:

- Bai et al. (“Efficient Calibration of Multi-Agent Market Simulators from Time Series with Bayesian Optimization,” 2021)
- Beikirch et al. (“Robust Mathematical Formulation and Probabilistic Description of Agent-Based Computational Economic Market Models,” 2021)
- Chabi-Yo and Loudis (“A Decomposition of Conditional Risk Premia and Implications for Representative Agent Models,” 2021)
- Clark et al. (“Test case generation for agent-based models: A systematic literature review,” 2021)
- De Gennaro Aquino et al. (“Portfolio Selection With Exploration of New Investment Opportunities,” 2021)
- Foramitti (“AgentPy: A package for agent-based modeling in Python,” 2021)
- Esmaeili et al. (“HAMLET: A Hierarchical Agent-based Machine Learning Platform,” 2021)
- Harwick (“Helipad: A Framework for Agent-Based Modeling in Python,” 2021)
- Istrate (“Models we Can Trust: Toward a Systematic Discipline of (Agent-Based) Model Interpretation and Validation,” 2021)
- Jericevich et al. (“Calibrating an adaptive Farmer-Joshi agent-based model for financial markets,” 2021)
- Karaš and Serwatka (“Strong-hand conjecture: agent-based numerical simulation,” 2021)
- Lindner et al. (“Learning What To Do by Simulating the Past,” 2021)
- Lux (“Can heterogeneous agent models explain the alleged mispricing of the S&P 500?” 2021)

Niemann et al. (“Data-driven model reduction of agent-based systems using the Koopman generator,” 2021)  
 Papoudakis et al. (“Agent Modelling under Partial Observability for Deep Reinforcement Learning,” 2021)  
 Perumal and van Zyl (“Surrogate Assisted Methods for the Parameterisation of Agent-Based Models,” 2020)  
 Sharma et al. (“A constraint-satisfaction agent-based model for the macroeconomy,” 2020)  
 Shi et al. (“Pyramid scheme in stock market: a kind of financial market simulation,” 2021)  
 Schuderer et al. (“Sim-Env: Decoupling OpenAI Gym Environments from Simulation Models,” 2021)  
 Sharma et al. (“V- U-, L- or W- shaped economic recovery after Covid-19: Insights from an Agent Based Model,” 2021)  
 ter Ellen et al. (“Comparing behavioural heterogeneity across asset classes,” 2021)  
 Thiem et al. (“Global and Local Reduced Models for Interacting, Heterogeneous Agents,” 2021)  
 Vandin et al. (“Automated and Distributed Statistical Analysis of Economic Agent-Based Models,” 2021)  
 Westphal and Sornette (“How Market Intervention Can Prevent Bubbles and Crashes,” 2020)

### 2.2.18 Validating scenarios and augmented datasets

List of references:

Alaa et al. (“How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models,” 2022)  
 Bahrpeyma et al. (“A Methodology for Validating Diversity in Synthetic Time Series Generation,” 2021)  
 Cramer et al. (“Validation Methods for Energy Time Series Scenarios from Deep Generative Models,” 2021)  
 De Meer et al. (“Tackling the exponential scaling of signature-based GANs for high-dimensional financial time series generation,” 2021)  
 Ding et al. (“Modeling Extreme Events in Time Series Prediction,” 2019)  
 Franco-Pedroso et al. (“The ETS challenges: a machine learning approach to the evaluation of simulated financial time series for improving generation processes,” 2019)  
 Franco-Pedroso et al. (“Generating Virtual Scenarios of Multivariate Financial Data for Quantitative Trading Applications,” 2019)  
 Istrate (“Models we Can Trust: Toward a Systematic Discipline of (Agent-Based) Model Interpretation and Validation,” 2021)  
 Lawrence et al. (“Data Generating Process to Evaluate Causal Discovery Techniques for Time Series Data,” 2021)  
 Ni et al. (“Conditional Sig-Wasserstein GANs for Time Series Generation,” 2020)  
 Snow (“DataGene: A Framework for Dataset Similarity,” 2020)

### 2.2.19 Software implementations and frameworks

List of references:

Foramitti (“AgentPy: A package for agent-based modeling in Python,” 2021)  
 Esmaeili et al. (“HAMLET: A Hierarchical Agent-based Machine Learning Platform,” 2021)  
 Harwick (“Helipad: A Framework for Agent-Based Modeling in Python,” 2021)  
 Hu et al. (“Multiple Imputation and Synthetic Data Generation with the R package NPBatesImputeCat,” 2021)  
 Jiang et al. (“A set of efficient methods to generate high-dimensional binary data with specified correlation structures,” 2021)  
 Kang et al. (“GRATIS: GeneRAting Time Series with diverse and controllable characteristics,” 2020)  
 Li et al. (“SynC: A Copula based Framework for Generating Synthetic Data from Aggregated Sources,” 2020)  
 Lin et al. (“Generating High-fidelity, Synthetic Time Series Datasets with DoppelGANger,” 2019)  
 Meyer and Nagler (“Synthia: multidimensional synthetic data generation in Python,” 2021)  
 Pesenti et al. (“Scenario weights for importance measurement (SWIM) - an R package for sensitivity analysis,” 2020)  
 Siebert et al. (“A systematic review of Python packages for time series analysis,” 2021)  
 Snow (“DeltaPy: A Framework for Tabular Data Augmentation in Python,” 2020)  
 Tadayon and Pottie (“tsBNgen: A Python Library to Generate Time Series Data from an Arbitrary Dynamic Bayesian Network Structure,” 2020)  
 Vinod (“Constructing Scenarios of Time Heterogeneous Series for Stress Testing,” 2012)  
 Vinod (“R Package GeneralCorr Functions for Portfolio Choice,” 2021)



Wang et al. (“VEGA: Towards an End-to-End Configurable AutoML Pipeline,” 2020)

## References

- Al Janabi, M. A. M., Arreola Hernandez, J., Berger, T., and Nguyen, D. K. (2017). “Multivariate dependence and portfolio optimization algorithms under illiquid market scenarios.” In: *European Journal of Operational Research* 259(3), pp. 1121–1131.

An analytical framework for Liquidity-adjusted Value-at-Risk is introduced. We develop a model for optimizing structured portfolios with liquidity constraints. Dynamic copula is used to gauge the multivariate dependence between assets. Our approach gives better efficient frontiers than competing optimization models. We propose a model for optimizing structured portfolios with liquidity-adjusted Value-at-Risk (LVaR) constraints, whereby linear correlations between assets are replaced by the multivariate nonlinear dependence structure based on Dynamic conditional correlation t-copula modeling. Our portfolio optimization algorithm minimizes the LVaR function under adverse market circumstances and multiple operational and financial constraints. When considering a diversified portfolio of international stock and commodity market indices under multiple realistic portfolio optimization scenarios, the obtained results consistently show the superiority of our approach, relative to other competing portfolio strategies including the minimum-variance, risk-parity and equally weighted portfolio allocations.

- Alaa, A., Van Breugel, B., Saveliev, E. S., and van der Schaar, M. (2022). “How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models.” In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, pp. 290–306.

Devising domain- and model-agnostic evaluation metrics for generative models is an important and as yet unresolved problem. Most existing metrics, which were tailored solely to the image synthesis setup, exhibit a limited capacity for diagnosing the different modes of failure of generative models across broader application domains. In this paper, we introduce a 3-dimensional evaluation metric, ( $\alpha$ -Precision,  $\beta$ -Recall, Authenticity), that characterizes the fidelity, diversity and generalization performance of any generative model in a domain-agnostic fashion. Our metric unifies statistical divergence measures with precision-recall analysis, enabling sample- and distribution-level diagnoses of model fidelity and diversity. We introduce generalization as an additional, independent dimension (to the fidelity-diversity trade-off) that quantifies the extent to which a model copies training data—a crucial performance indicator when modeling sensitive data with requirements on privacy. The three metric components correspond to (interpretable) probabilistic quantities, and are estimated via sample-level binary classification. The sample-level nature of our metric inspires a novel use case which we call model auditing, wherein we judge the quality of individual samples generated by a (black-box) model, discarding low-quality samples and hence improving the overall model performance in a post-hoc manner.

- Albanese, C., Crepey, S., and Iabichino, S. (2020). “Reverse Stress Testing.” In: *SSRN e-Print*.

This article outlines a framework for the analysis of extreme events based on forward-looking reverse stress testing. We carry out a portfolio simulation and identify stress scenarios which are critical for bank solvency as the ones contributing the most to cost of capital, as expressed by KVA scenario differentials. Applications include model validation, trading limits, model risk management and hedging. A pricing model is invalid if it breaks on a path leading to stress conditions, causing alpha leaks that go undetected in market risk models such as value-at-risk (VaR), expected shortfall and stressed VaR. Trading limits are best predicated on incremental cost of capital and model risk capital can be assessed by computing cost of capital with Bayesian averages. Stress scenarios also have the potential to suggest risk-reducing hedges.

- Alexander, C. and Ledermann, D. (2012). “ROM Simulation: Applications to Stress Testing and VaR.” In: *SSRN e-Print*.

Most banks employ historical simulation for Value-at-Risk (VaR) calculations, where VaR is computed from a lower quantile of a forecast distribution for the portfolio’s profit and loss (P and L) that is constructed from a single, multivariate historical sample on the portfolio’s risk factors. The implicit assumption is that history will repeat itself for certain over the forecast horizon. Until now, the only alternative is to assume the historical sample is generated by a multivariate, parametric risk factor distribution and (except in special cases where an analytic solution is available) to simulate P and L via Monte Carlo (MC). This paper introduces a methodology that encompasses historical and MC VaR as special cases, which is much faster than MC simulation and which avoids the single-sample bias of historical simulation. Random orthogonal matrix (ROM) simulation is a fast matrix-based simulation method that applies directly to an historical sample, or to a parametric distribution. Each simulation matches the first four multivariate sample moments to those of the observed sample, or of

the target distribution. Stressed VaR is typically computed from an historical sample using the Duffie-Pan methodology, whereby the sample is transformed to have a stressed covariance matrix. ROM simulation extends this methodology to generate very large samples, which furthermore have stressed values for the first four multivariate moments values.

Alokley, S. A. and Albarrak, M. S. (2020). “Clustering of Extremes in Financial Returns: A Study of Developed and Emerging Markets.” In: *Journal of Risk and Financial Management* 13(7), p. 141.

This paper investigates the clustering or dependency of extremes in financial returns by estimating the extremal index value, in which smaller values of the extremal index correspond to more clustering. We apply the interval estimator method to determine the extremal index for a range of threshold values in the developed and emerging markets from 2007-2017. The indices we used to represent developed markets are from France, Germany, Italy, Japan, USA, UK, Spain, and Sweden. For the emerging markets, we use indices from China, Brazil, India, Malaysia, Russia, Saudi Arabia, and Portugal. The results show that clustering occurs in the emerging and developed markets under several threshold values. This study will shed light on the dependency structure of financial returns data and the proprieties of the extremes returns. Moreover, understanding clustering of extremes in these markets can help investors reduce the exposure to extreme financial events, such as the financial crisis.

Amengual, D. and Sentana, E. (2020). “Is a Normal Copula the Right Copula?” In: *Journal of Business & Economic Statistics* 38(2), pp. 350–366.

We derive computationally simple and intuitive expressions for score tests of Gaussian copulas against generalized hyperbolic alternatives, including symmetric and asymmetric Student t, and many other examples. We decompose our tests into third and fourth moment components, and obtain one-sided Likelihood Ratio analogs, whose standard asymptotic distribution we provide. Our Monte Carlo exercises confirm the reliable size of parametric bootstrap versions of our tests, and their substantial power gains over alternative procedures. In an empirical application to CRSP stocks, we find that short-term reversals and momentum effects are better captured by non-Gaussian copulas, whose parameters we estimate by indirect inference. Supplementary materials for this article are available online.

Ardia, D. and Bluteau, K. (2017). “Stress-Testing With Parametric Models and Fully Flexible Probabilities.” In: *Wilmott* 2017(87), pp. 52–55.

We propose a simple methodology to simulate scenarios from a parametric risk model while accounting for stress-test views via fully flexible probabilities.

Ardia, D. and Meucci, A. (2015). “Stress testing in non-normal markets via entropy pooling.” In: *Risk*.

The combination of trading signals, or general views on the market, within a prior risk model to compute an optimal allocation that incorporates these views is one of the main challenges in quantitative portfolio construction. Similarly, embedding stress tests in a risk model in a statistically sound way is key to a healthy risk management process.

Arora, R., Gao, R., and Tompaidis, S. (2021). “Extreme Yet Plausible: Choosing Scenarios to Stress Test Financial Institutions.” In: *SSRN e-Print*.

Since the 2008 Financial Crisis, stress tests based on extreme-yet-plausible scenarios have become a preferred method of assessing risk for large financial institutions, yet scenario choice has largely been ad-hoc. We propose a principled methodology to choose scenarios by minimizing the estimation error of the risk of the profit and loss (P/L) distribution of a financial institution. We consider three separate cases: 1) a parametric case, when the P/L depends linearly on risk factors that are assumed to be elliptically distributed; 2) a distributionally robust case, where the P/L depends linearly on risk factors whose distribution is assumed to lie within an uncertainty set based on the Wasserstein metric; 3) a non-linear case, when the P/L is a non-linear function of the risk factors. Our methodology connects stress testing with the area of Design of Experiments with a risk-based design objective. We illustrate the advantages of our framework by comparing the accuracy of the CVaR estimate achieved with stress scenarios chosen under our framework to the accuracy achieved by stress scenarios chosen in 2019 by the Commodity Futures Trading Commission to stress test central counterparties.

Asch, A., Brady, E., Gallardo, H., Hood, J., Chu, B., and Farazmand, M. (2022). “Model-assisted deep learning of rare extreme events from partial observations.” In: *arXiv e-Print*.

To predict rare extreme events using deep neural networks, one encounters the so-called small data problem because even long-term observations often contain few extreme events. Here, we investigate a model-assisted framework where the training data is obtained from numerical simulations, as opposed to observations, with adequate samples from extreme events. However, to ensure the trained networks are applicable in practice, the training is not performed on the full simulation data; instead we only use a small subset of observable quantities

which can be measured in practice. We investigate the feasibility of this model-assisted framework on three different dynamical systems (Rossler attractor, FitzHugh–Nagumo model, and a turbulent fluid flow) and three different deep neural network architectures (feedforward, long short-term memory, and reservoir computing). In each case, we study the prediction accuracy, robustness to noise, reproducibility under repeated training, and sensitivity to the type of input data. In particular, we find long short-term memory networks to be most robust to noise and to yield relatively accurate predictions, while requiring minimal fine-tuning of the hyperparameters.

- Assefa, S., Dervovic, D., Mahfouz, M., Balch, T., Reddy, P., and Veloso, M. (2019). “[Generating Synthetic Data in Finance: Opportunities, Challenges and Pitfalls.](#)” In: *NeurIPS’19 Workshop on Robust AI in Financial Services*. Financial services generate a huge volume of data that is extremely complex and varied. These datasets are often stored in silos within organisations for various reasons, including but not limited to, regulatory requirements and business needs. As a result, data sharing within different lines of business as well as outside of the organisation (e.g. to the research community) is severely limited. It is therefore critical to investigate methods for synthesising financial datasets that follow the same properties of the real data while respecting the need for privacy of the parties involved in a particular dataset. This introductory paper aims to highlight the growing need for effective synthetic data generation in the financial domain. We highlight three main areas of focus for the academic community: 1) Generating realistic synthetic datasets. 2) Measuring the similarities between real and generated datasets 3) Ensuring the generative process satisfies any privacy constraints. Although these challenges are also present in other domains, the extra regulatory and privacy requirements add another dimension of complexity and offer a unique opportunity to study the topic in financial services. Finally, we aim to develop a shared vocabulary and context for generating synthetic financial data using two types of financial datasets as examples.
- Aste, T. (2021). “[Stress testing and systemic risk measures using multivariate conditional probability.](#)” In: *arXiv e-Print*.

The multivariate conditional probability distribution quantifies the effects of a set of variables onto the statistical properties of another set of variables. In the study of systemic risk in the financial system, the multivariate conditional probability distribution can be used for stress-testing by quantifying the propagation of losses from a set of ‘stressing’ variables to another set of ‘stressed’ variables. Here it is described how to compute such conditional probability distributions for the vast family of multivariate elliptical distributions, which includes the multivariate Student-t and the multivariate Normal distributions. Simple measures of stress impact and systemic risk are also proposed. An application to the US equity market illustrates the potentials of this approach.

- Baes, M. and Schaanning, E. (2020). “[Reverse Stress Testing: Scenario Design for Macroprudential Stress Tests.](#)” In: *SSRN e-Print*.

We propose a systematic algorithmic reverse-stress testing methodology to create “worst case” scenarios for regulatory stress tests by accounting for losses that arise from distressed portfolio liquidations. First, we derive the optimal bank response for any given shock. Then, we introduce an algorithm which systematically generates scenarios that exploit the key vulnerabilities in banks’ portfolio holdings and thus maximize contagion despite banks’ optimal response to the shock. We apply our methodology to data of the 2016 European Banking Authority (EBA) stress test, and design worst case scenarios for the portfolio holdings of European banks at the time. Using spectral clustering techniques, we group 10’000 worst-case scenarios into twelve geographically concentrated families. Our results show that even though there is a wide range of different scenarios within these twelve families, each cluster tends to affect the same banks. An “Anna Karenina” principle of stress testing emerges: Not all stressful scenarios are alike, but every stressful scenario stresses the same banks. These findings suggest that the precise specification of a scenario is not of primal importance as long as the most vulnerable banks are targeted and sufficiently stressed. Finally, our methodology can be used to uncover the weakest links in the financial system and thereby focus supervisory attention on these, thus building a bridge between macro-prudential and micro-prudential stress tests.

- Bahrpeyma, F., Roantree, M., Cappellari, P., Scriney, M., and McCarren, A. (2021). “[A Methodology for Validating Diversity in Synthetic Time Series Generation.](#)” In: *MethodsX* 8, p. 101459.

In order for researchers to deliver robust evaluations of time series models, it often requires high volumes of data to ensure the appropriate level of rigor in testing. However, for many researchers, the lack of time series presents a barrier to a deeper evaluation. While researchers have developed and used synthetic datasets, the development of this data requires a methodological approach to testing the entire dataset against a set of metrics which capture the diversity of the dataset. Unless researchers are confident that their test datasets display a broad set of time series characteristics, it may favor one type of predictive model over another. This can have the effect of undermining the evaluation of new predictive methods. In this paper, we present a new approach to generating

and evaluating a high number of time series data. The construction algorithm and validation framework are described in detail, together with an analysis of the level of diversity present in the synthetic dataset.

- Bai, Y., Lam, H., Vyetrenko, S., and Balch, T. (2021). “Efficient Calibration of Multi-Agent Market Simulators from Time Series with Bayesian Optimization.” In: *arXiv e-Print*.

Multi-agent market simulation is commonly used to create an environment for downstream machine learning or reinforcement learning tasks, such as training or testing trading strategies before deploying them to real-time trading. In electronic trading markets only the price or volume time series, that result from interaction of multiple market participants, are typically directly observable. Therefore, multi-agent market environments need to be calibrated so that the time series that result from interaction of simulated agents resemble historical – which amounts to solving a highly complex large-scale optimization problem. In this paper, we propose a simple and efficient framework for calibrating multi-agent market simulator parameters from historical time series observations. First, we consider a novel concept of eligibility set to bypass the potential non-identifiability issue. Second, we generalize the two-sample Kolmogorov-Smirnov (K-S) test with Bonferroni correction to test the similarity between two high-dimensional time series distributions, which gives a simple yet effective distance metric between the time series sample sets. Third, we suggest using Bayesian optimization (BO) and trust-region BO (TuRBO) to minimize the aforementioned distance metric. Finally, we demonstrate the efficiency of our framework using numerical experiments.

- 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.

- Barra Montevecchi, J. A., Silva Costa, R. F. da, Pinho, A. F. de, and Carvalho Miranda, R. de (2017). “A simulation-based approach to perform economic evaluation scenarios.” In: *Journal of Simulation* 11(2), pp. 185–192.

The interest in the joint use of discrete event simulation (DES) with design of experiments (DOE), activity-based costing (ABC) and net present value (NPV) in order to aid in decision making in manufacturing systems has been on the rise in the past decade. Traditionally, the literature has presented research combining simulation with one or two of the mentioned techniques. In this paper, all four mentioned techniques are integrated into one methodology in order to economically evaluate simulation model scenarios for manufacturing systems. Scenarios are simulated and analysed through the application of ABC and DOE. Then, the significant scenarios are analysed through NPV. Through application of the methodology in a production cell, it is shown that it is viable to utilize these four techniques together. This research contributes to current literature through the development of a methodology that combines DES, ABC, DOE and NPV, in a decision-making structure.

- Bass, R., Gallagher, K., Ratcliffe, R., and Shah, S. (2018). “Factor Performance Across Market-Driven Scenarios.” In: *SSRN e-Print*.

We develop a methodology to model factor returns in scenarios which are yet to occur or do not correspond to past market conditions. We infer factor performance conditional on hypothetical market-driven scenarios, which are determined by a parsimonious number of underlying policy shocks. We derive regime-dependent corresponding security and asset class returns, which are functions of factors and policy drivers. The analysis is helpful in constructing robust portfolios designed for more stable performance across various scenarios, and can aid in framing and communicating portfolio risks.

- Beikirch, M., Cramer, S., Frank, M., Otte, P., Pabich, E., and Trimborn, T. (2021). “Robust Mathematical For-

mulation and Probabilistic Description of Agent-Based Computational Economic Market Models.” In: *arXiv e-Print*.

In science and especially in economics, agent-based modeling has become a widely used modeling approach. These models are often formulated as a large system of difference equations. In this study, we discuss two aspects, numerical modeling and the probabilistic description for two agent-based computational economic market models: the Levy-Levy-Solomon model and the Franke-Westerhoff model. We derive time-continuous formulations of both models, and in particular we discuss the impact of the time-scaling on the model behavior for the Levy-Levy-Solomon model. For the Franke-Westerhoff model, we proof that a constraint required in the original model is not necessary for stability of the time-continuous model. It is shown that a semi-implicit discretization of the time-continuous system preserves this unconditional stability. In addition, this semi-implicit discretization can be computed at cost comparable to the original model. Furthermore, we discuss possible probabilistic descriptions of time continuous agent-based computational economic market models. Especially, we present the potential advantages of kinetic theory in order to derive mesoscopic descriptions of agent-based models. Exemplified, we show two probabilistic descriptions of the Levy-Levy-Solomon and Franke-Westerhoff model.

Benali, F., Bodenes, D., Labroche, N., and de Runz, C. (2021). “MTCopula: Synthetic Complex Data Generation Using Copula.” In: *HAL Archives Ouvertes*.

Nowadays, marketing strategies are data-driven, and their quality depends significantly on the quality and quantity of available data. As it is not always possible to access this data, there is a need for synthetic data generation. Most of the existing techniques work well for low-dimensional data and may fail to capture complex dependencies between data dimensions. Moreover, the tedious task of identifying the right combination of models and their respective parameters is still an open problem. In this paper, we present MTCopula, a novel approach for synthetic complex data generation based on Copula functions. MTCopula is a flexible and extendable solution that automatically chooses the best Copula model, between Gaussian Copula and T-Copula models, and the best-fitted marginals to catch the data complexity. It relies on Maximum Likelihood Estimation to fit the possible marginal distribution models and introduces Akaike Information Criterion to choose both the best marginals and Copula models, thus removing the need for a tedious manual exploration of their possible combinations. Comparisons with state-of-art synthetic data generators on a real use case private dataset, called AdWanted, and literature datasets show that our approach preserves better the variable behaviors and the dependencies between variables in the generated synthetic datasets.

Bertsimas, D. and Van Parys, B. (2017). “Bootstrap Robust Prescriptive Analytics.” In: *arXiv e-Print*.

We address the problem of prescribing an optimal decision in a framework where its cost depends on uncertain problem parameters  $Y$  that need to be learned from data. Earlier work by Bertsimas and Kallus (2014) transforms classical machine learning methods that merely predict  $Y$  from supervised training data  $[(x_1, y_1), \dots, (x_n, y_n)]$  into prescriptive methods taking optimal decisions specific to a particular covariate context  $X = x$ . Their prescriptive methods factor in additional observed contextual information on a potentially large number of covariates  $X = x$  to take context specific actions  $z(x)$  which are superior to any static decision  $z$ . Any naive use of limited training data may, however, lead to gullible decisions over-calibrated to one particular data set. In this paper, we borrow ideas from distributionally robust optimization and the statistical bootstrap of Efron (1982) to propose two novel prescriptive methods based on (nw) Nadaraya-Watson and (nn) nearest-neighbors learning which safeguard against overfitting and lead to improved out-of-sample performance. Both resulting robust prescriptive methods reduce to tractable convex optimization problems and enjoy a limited disappointment on bootstrap data. We illustrate the data-driven decision-making framework and our novel robustness notion on a small news vendor problem as well as a small portfolio allocation problem.

Bhatia, S., Jain, A., and Hooi, B. (2021). “ExGAN: Adversarial Generation of Extreme Samples.” In: *arXiv e-Print*.

Mitigating the risk arising from extreme events is a fundamental goal with many applications, such as the modelling of natural disasters, financial crashes, epidemics, and many others. To manage this risk, a vital step is to be able to understand or generate a wide range of extreme scenarios. Existing approaches based on Generative Adversarial Networks (GANs) excel at generating realistic samples, but seek to generate typical samples, rather than extreme samples. Hence, in this work, we propose ExGAN, a GAN-based approach to generate realistic and extreme samples. To model the extremes of the training distribution in a principled way, our work draws from Extreme Value Theory (EVT), a probabilistic approach for modelling the extreme tails of distributions. For practical utility, our framework allows the user to specify both the desired extremeness measure, as well as the desired extremeness probability they wish to sample at. Experiments on real US Precipitation data show



that our method generates realistic samples, based on visual inspection and quantitative measures, in an efficient manner. Moreover, generating increasingly extreme examples using ExGAN can be done in constant time (with respect to the extremeness probability), as opposed to the exponential time required by the baseline approach.

Bhattacharai, B., Baek, S., Bodur, R., and Kim, T.-K. (2020). “Sampling Strategies for GAN Synthetic Data.” In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Generative Adversarial Networks (GANs) have been used widely to generate large volumes of synthetic data. This data is being utilized for augmenting with real examples in order to train deep Convolutional Neural Networks (CNNs). Studies have shown that the generated examples lack sufficient realism to train deep CNNs and are poor in diversity. Unlike previous studies of randomly augmenting the synthetic data with real data, we present our simple, effective and easy to implement synthetic data sampling methods to train deep CNNs more efficiently and accurately. To this end, we propose to maximally utilize the parameters learned during training of the GAN itself. These include discriminator’s realism confidence score and the confidence on the target label of the synthetic data. In addition to this, we explore reinforcement learning (RL) to automatically search a subset of meaningful synthetic examples from a large pool of GAN synthetic data. We evaluate our method on two challenging face attribute classification data sets viz. AffectNet and CelebA. Our extensive experiments clearly demonstrate the need of sampling synthetic data before augmentation, which also improves the performance of one of the state-of-the-art deep CNNs in vitro.

Bianchi, M. L., Stoyanov, S. V., Tassinari, G. L., Fabozzi, F. J., and Focardi, S. M. (2019). *Handbook of Heavy-Tailed Distributions in Asset Management and Risk Management*. World Scientific. 600 pp.

The study of heavy-tailed distributions allows researchers to represent phenomena that occasionally exhibit very large deviations from the mean. The dynamics underlying these phenomena is an interesting theoretical subject, but the study of their statistical properties is in itself a very useful endeavor from the point of view of managing assets and controlling risk. In this book, the authors are primarily concerned with the statistical properties of heavy-tailed distributions and with the processes that exhibit jumps. A detailed overview with a Matlab implementation of heavy-tailed models applied in asset management and risk managements is presented. The book is not intended as a theoretical treatise on probability or statistics, but as a tool to understand the main concepts regarding heavy-tailed random variables and processes as applied to real-world applications in finance. Accordingly, the authors review approaches and methodologies whose realization will be useful for developing new methods for forecasting of financial variables where extreme events are not treated as anomalies, but as intrinsic parts of the economic process.

Bilgili, M. S., Ferconi, M., and Ulitsky, A. (2017). “Stress Hedging in Portfolio Construction.” In: *Risk*.

Scenario stress testing is a useful and increasingly popular approach to assess portfolio performance under different market conditions. In this paper we focus on how to incorporate stress scenario information directly in portfolio construction as additional constraints to control for potential losses and risks. To broaden the applicability of stress testing we propose robust constrained optimization approach for handling uncertainty in scenario parameters. An example of the – Oil-Crisis– event is used as a numerical illustration.

Bolger, F. and Wright, G. (2017). “Use of expert knowledge to anticipate the future: Issues, analysis and directions.” In: *International Journal of Forecasting* 33(1), pp. 230–243.

Unless an anticipation problem is routine and short-term, and objective data are plentiful, expert judgment will be needed. Risk assessment is analogous to anticipating the future, in that models need to be developed and applied to data. Since objective data are often scanty, expert knowledge elicitation (EKE) techniques have been developed for risk assessment that allow models to be developed and parametrized using expert judgment with minimal cognitive and social biases. Here, we conceptualize how EKE can be developed and applied to support anticipation of the future. Accordingly, we begin by defining EKE as a complete process, which involves considering experts as a source of data, and comprises various methods for ensuring the quality of this data, including selecting the best experts, training experts in the normative aspects of anticipation, and combining judgments from several experts, as well as eliciting unbiased estimates and constructs from experts. We detail various aspects of the papers that constitute this special issue and analyse them in terms of the stages of the EKE future-anticipation process that they address. We also identify the remaining gaps in our knowledge. Our conceptualization of EKE with the aim of supporting anticipation of the future is compared and contrasted with the extant research on judgmental forecasting.

Bond-Taylor, S., Leach, A., Long, Y., and Willcocks, C. G. (2022). “Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models.” In: *arXiv e-Print*.



Deep generative modelling is a class of techniques that train deep neural networks to model the distribution of training samples. Research has fragmented into various interconnected approaches, each of which making trade-offs including run-time, diversity, and architectural restrictions. In particular, this compendium covers energy-based models, variational autoencoders, generative adversarial networks, autoregressive models, normalizing flows, in addition to numerous hybrid approaches. These techniques are drawn under a single cohesive framework, comparing and contrasting to explain the premises behind each, while reviewing current state-of-the-art advances and implementations.

Boubaker, S. and Nguyen, D. K. (2019). *Handbook of Global Financial Markets*. World Scientific. 828 pp.

The objective of this handbook is to provide the readers with insights about current dynamics and future potential transformations of global financial markets. We intend to focus on four main areas: Dynamics of Financial Markets; Financial Uncertainty and Volatility; Market Linkages and Spillover Effects; and Extreme Events and Financial Transformations and address the following critical issues, but not limited to: market integration and its implications; crisis risk assessment and contagion effects; financial uncertainty and volatility; role of emerging financial markets in the global economy; role of complex dynamics of economic and financial systems; market linkages, asset valuation and risk management; exchange rate volatility and firm-level exposure; financial effects of economic, political and social risks; link between financial development and economic growth; country risks; and sovereign debt markets.

Breuer, T. and Summer, M. (2020). “Systematic stress tests on public data.” In: *Journal of Banking & Finance* 118, pp. 105886+.

For a given set of banks, how big can losses in bad economic or financial scenarios possibly get, and what are these bad scenarios? These are the two central questions of stress tests for banks and the banking system. Current stress tests select stress scenarios in a way which might leave aside many dangerous scenarios and thus create an illusion of safety; and which might consider highly implausible scenarios and thus trigger a false alarm. We show how to select scenarios systematically for a banking system in a context of multiple credit exposures. We demonstrate the application of our method in an example on the Spanish and Italian residential real estate exposures of European banks. Compared to the EBA 2016 stress test our method produces scenarios which are equally plausible as the EBA stress scenario but yield considerably worse system wide losses.

Buczak, A. L., Baugher, B. D., Martin, C. S., Keiley-Listermann, M. W., Howard, J., Parrish, N. H., Stalick, A. Q., Berman, D. S., and Dredze, M. H. (2022). “Crystal Cube: Forecasting Disruptive Events.” In: *Applied Artificial Intelligence* 36(1) (2001179), pp. 1–24.

Disruptive events within a country can have global repercussions, creating a need for the anticipation and planning of these events. Crystal Cube (CC) is a novel approach to forecasting disruptive political events at least one month into the future. The system uses a recurrent neural network and a novel measure of event similarity between past and current events. We also introduce the innovative Thermometer of Irregular Leadership Change (ILC). We present an evaluation of CC in predicting ILC for 167 countries and show promising results in forecasting events one to twelve months in advance. We compare CC results with results using a random forest as well as previous work.

Buehler, H., Gonon, L., Teichmann, J., and Wood, B. (2019). “Deep hedging.” In: *Quantitative Finance* 19 (8), pp. 1271–1291.

We present a framework for hedging a portfolio of derivatives in the presence of market frictions such as transaction costs, liquidity constraints or risk limits using modern deep reinforcement machine learning methods. We discuss how standard reinforcement learning methods can be applied to non-linear reward structures, i.e. in our case convex risk measures. As a general contribution to the use of deep learning for stochastic processes, we also show in Section 4 that the set of constrained trading strategies used by our algorithm is large enough to eps-approximate any optimal solution. Our algorithm can be implemented efficiently even in high-dimensional situations using modern machine learning tools. Its structure does not depend on specific market dynamics, and generalizes across hedging instruments including the use of liquid derivatives. Its computational performance is largely invariant in the size of the portfolio as it depends mainly on the number of hedging instruments available. We illustrate our approach by an experiment on the SandP500 index and by showing the effect on hedging under transaction costs in a synthetic market driven by the Heston model, where we outperform the standard -market solution.

Buehler, H., Horvath, B., Lyons, T., Arribas, I. P., and Wood, B. (2020). “A Data-driven Market Simulator for Small Data Environments.” In: *arXiv e-Print*.

Neural network based data-driven market simulation unveils a new and flexible way of modelling financial time series without imposing assumptions on the underlying stochastic dynamics. Though in this sense generative market simulation is model-free, the concrete modelling choices are nevertheless decisive for the features of the simulated paths. We give a brief overview of currently used generative modelling approaches and performance evaluation metrics for financial time series, and address some of the challenges to achieve good results in the latter. We also contrast some classical approaches of market simulation with simulation based on generative modelling and highlight some advantages and pitfalls of the new approach. While most generative models tend to rely on large amounts of training data, we present here a generative model that works reliably in environments where the amount of available training data is notoriously small. Furthermore, we show how a rough paths perspective combined with a parsimonious Variational Autoencoder framework provides a powerful way for encoding and evaluating financial time series in such environments where available training data is scarce. Finally, we also propose a suitable performance evaluation metric for financial time series and discuss some connections of our Market Generator to deep hedging.

Buehler, H., Horvath, B., Lyons, T., Arribas, I. P., and Wood, B. (2021). “Generating financial markets with signatures.” In: *Risk*.

While most generative models tend to rely on large amounts of training data, here Hans Buehler et al present a generative model that works reliably even in environments where the amount of available training data is small, irregularly paced or oscillatory. They show how a rough paths-based feature map encoded by the signature of the path outperforms returns-based market generation both numerically and from a theoretical point of view. Finally, they propose a suitable performance evaluation metric for financial time series and discuss some connections between their signature-based market generator and deep hedging.

Cai, M.-L., Chen, Z.-H., Li, S.-P., Xiong, X., Zhang, W., Yang, M.-Y., and Ren, F. (2022). “New volatility evolution model after extreme events.” In: *arXiv e-Print*.

In this paper, we propose a new dynamical model to study the two-stage volatility evolution of stock market index after extreme events, and find that the volatility after extreme events follows a stretched exponential decay in the initial stage and becomes a power law decay at later times by using high frequency minute data. Empirical study of the evolutionary behaviors of volatility after endogenous and exogenous events further demonstrates the descriptive power of our new model. To further explore the underlying mechanisms of volatility evolution, we introduce the sequential arrival of information hypothesis (SAIH) and the mixture of distribution hypothesis (MDH) to test the two-stage assumption, and find that investors transform from the uninformed state to the informed state in the first stage and informed investors subsequently dominate in the second stage. The testing results offer a supporting explanation for the validity of our new model and the fitted values of relevant parameters.

Cao, H. and Guo, X. (2021). “Generative Adversarial Network: Some Analytical Perspectives.” In: *arXiv e-Print*.

Ever since its debut, generative adversarial networks (GANs) have attracted tremendous amount of attention. Over the past years, different variations of GANs models have been developed and tailored to different applications in practice. Meanwhile, some issues regarding the performance and training of GANs have been noticed and investigated from various theoretical perspectives. This subchapter will start from an introduction of GANs from an analytical perspective, then move on the training of GANs via SDE approximations and finally discuss some applications of GANs in computing high dimensional MFGs as well as tackling mathematical finance problems.

Carrillo, J. A., Nieto, M., Velez, J. F., and Velez, D. (2021). “A New Machine Learning Forecasting Algorithm Based on Bivariate Copula Functions.” In: *Forecasting* 3(2), pp. 355–376.

A novel forecasting method based on copula functions is proposed. It consists of an iterative algorithm in which a dependent variable is decomposed as a sum of error terms, where each one of them is estimated identifying the input variable which best “copulate” with it. The method has been tested over popular reference datasets, achieving competitive results in comparison with other well-known machine learning techniques.

Cavaliere, G., Nielsen, H. B., and Rahbek, A. (2020). “An Introduction to Bootstrap Theory in Time Series Econometrics.” In: *SSRN e-Print*.

This article provides an introduction to methods and challenges underlying application of the bootstrap in econometric modelling of economic and financial time series. Validity, or asymptotic validity, of the bootstrap is discussed as this is a key element in deciding whether the bootstrap is applicable in empirical contexts. That is, as detailed here, bootstrap validity relies on regularity conditions, which need to be verified on a case-by-case basis. To fix ideas, asymptotic validity is discussed in terms of the leading example of bootstrap-based

hypothesis testing in the well-known first order auto-regressive model. In particular, bootstrap versions of classic convergence in probability and distribution, and hence of laws of large numbers and central limit theorems, are discussed as crucial ingredients to establish bootstrap validity. Regularity conditions and their implications for possible improvements in terms of (empirical) size and power for bootstrap-based testing, when compared to asymptotic testing, are illustrated by simulations. Following this, an overview of selected recent advances in the application of bootstrap methods in econometrics is also given.

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.

Chabi-Yo, F. and Loudis, J. (2021). “A Decomposition of Conditional Risk Premia and Implications for Representative Agent Models.” In: *SSRN e-Print*.

We develop a methodology to decompose the conditional market risk premium and risk premia on arbitrary moments of excess market returns into components related to contingent claims on down, up, and normal market returns. We call these components the downside, upside, and central risk premia. The decomposition does not depend on assumptions about investor preferences or the market return distribution, and can be computed in real time using a cross-section of option prices. The components’ contributions to total risk premia vary significantly over time and across investment horizon. Our risk premium decomposition offers a powerful tool for evaluating representative agent models in a conditional setting. We develop a related methodology to estimate analogous conditional decompositions implied by leading representative agent models, and compare these to data-implied decompositions. Although many representative agent models match the unconditional market risk premium, they generally do a poor job matching the downside, central, and upside risk premia both conditionally and unconditionally.

Chadebec, C. and Allasonniere, S. (2021). “Data Augmentation with Variational Autoencoders and Manifold Sampling.” In: *arXiv e-Print*.

We propose a new efficient way to sample from a Variational Autoencoder in the challenging low sample size setting. This method reveals particularly well suited to perform data augmentation in such a low data regime and is validated across various standard and real-life data sets. In particular, this scheme allows to greatly improve classification results on the OASIS database where balanced accuracy jumps from 80.7% for a classifier trained with the raw data to 88.6% when trained only with the synthetic data generated by our method. Such results were also observed on 3 standard data sets and with other classifiers. A code is available at [https://github.com/clementchadebec/Data\\_Augmentation\\_with\\_VAE-DALI](https://github.com/clementchadebec/Data_Augmentation_with_VAE-DALI).

Chalkis, A., Christoforou, E., Dalamagkas, T., and Emiris, I. Z. (2021). “Modeling of crisis periods in stock markets.” In: *arXiv e-Print*.

We exploit a recent computational framework to model and detect financial crises in stock markets, as well as shock events in cryptocurrency markets, which are characterized by a sudden or severe drop in prices. Our method manages to detect all past crises in the French industrial stock market starting with the crash of 1929, including financial crises after 1990 (e.g. dot-com bubble burst of 2000, stock market downturn of 2002), and all past crashes in the cryptocurrency market, namely in 2018, and also in 2020 due to covid-19. We leverage copulae clustering, based on the distance between probability distributions, in order to validate the reliability of the framework; we show that clusters contain copulae from similar market states such as normal states, or crises. Moreover, we propose a novel regression model that can detect successfully all past events using less than 10% of the information that the previous framework requires. We train our model by historical data on the

industry assets, and we are able to detect all past shock events in the cryptocurrency market. Our tools provide the essential components of our software framework that offers fast and reliable detection, or even prediction, of shock events in stock and cryptocurrency markets of hundreds of assets.

Chalongvorachai, T. and Woraratpanya, K. (2021). “A data generation framework for extremely rare case signals.” In: *Helyon* 7(8), e07687.

Unlike data augmentation, data generation for extremely rare cases is an approach that can spawn a significant number of high-quality samples based on very few original data. This could be useful in anomaly detection and classification tasks that have the limitation of publicly available datasets for research purposes. Though some other approaches have attempted to solve this problem, such as data augmentation techniques, there was nothing to ensure the characteristics of synthesized samples. Previously, we initiated a framework, called Data Augmentation and Generation for Anomalous Time-series Signals (DAGAT), that was in cooperation with important components: Data Augmentation, Variational Autoencoder (VAE), Data Picker (DP), Signal Fragment Assembler (SFA), and Quality Classifier (QC). And then, an upgraded framework, called An Advanced Data Generation for Anomalous Signals (ADGAS), was introduced to eliminate the limitations of DAGAT; those are uncontrollable outputs and the possibility of bad data included in a training set. By reforming DAGAT architecture, ADGAS achieves a better outcome of generated samples. Nonetheless, ADGAS could be improved through better SFA, DP, and QC. Hence, this paper proposed a Data Generation Framework for Extremely Rare Case Signals. The proposed framework is achievable in generating reliable data for various objectives. We challenged this framework by using the 1D-CNN to serve as the performance evaluator in multi-class anomalous classifications and using the water treatment and water distribution testbed (SWaT and WADI) as the real-world anomaly datasets. The result shows that it surpasses other baseline methods of anomaly data augmentation and data generation techniques.

Charte, D., Charte, F., Garcia, S., Jesus, M. J. del, and Herrera, F. (2018). “A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines.” In: *Information Fusion* 44, pp. 78–96.

Many of the existing machine learning algorithms, both supervised and unsupervised, depend on the quality of the input characteristics to generate a good model. The amount of these variables is also important, since performance tends to decline as the input dimensionality increases, hence the interest in using feature fusion techniques, able to produce feature sets that are more compact and higher level. A plethora of procedures to fuse original variables for producing new ones has been developed in the past decades. The most basic ones use linear combinations of the original variables, such as PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis), while others find manifold embeddings of lower dimensionality based on non-linear combinations, such as Isomap or LLE (Linear Locally Embedding) techniques. More recently, autoencoders (AEs) have emerged as an alternative to manifold learning for conducting nonlinear feature fusion. Dozens of AE models have been proposed lately, each with its own specific traits. Although many of them can be used to generate reduced feature sets through the fusion of the original ones, there also AEs designed with other applications in mind. The goal of this paper is to provide the reader with a broad view of what an AE is, how they are used for feature fusion, a taxonomy gathering a broad range of models, and how they relate to other classical techniques. In addition, a set of didactic guidelines on how to choose the proper AE for a given task is supplied, together with a discussion of the software tools available. Finally, two case studies illustrate the usage of AEs with datasets of handwritten digits and breast cancer.

Chen, C. Y.-H. and Nasekin, S. (2020). “Quantifying systemic risk with factor copulas.” In: *The European Journal of Finance* 26(18), pp. 1926–1947.

We propose a tail dependence based network approach to study systemic risk in a network of systemically important financial institutions (SIFIs). We utilize a flexible factor copula based method which allows to measure the level of extreme risk in a portfolio when dependence is driven by one or several factors. We identify the most “connected” SIFIs based on an eigenvector centrality approach applied to copula-implied dependence structures as “central” SIFIs. We then demonstrate that the level of systemic risk implied by such SIFIs chosen as conditioning factors in the factor copula setup exceeds that which is implied by non-central SIFIs in terms of portfolio Value-at-Risk and the portfolio return under stress. This study contributes to quantification and ranking of the systemic importance of SIFIs which is important for setting adequate capital requirements in particular and stability of financial markets in general.

Chen, D.-G. and Chen, J. D. (2017). *Monte-Carlo Simulation-Based Statistical Modeling*. Ed. by D.-G. Chen and J. D. Chen. ICSA book series in statistics. Singapore: Springer Singapore. 424 pp.

This book brings together expert researchers engaged in Monte-Carlo simulation-based statistical modeling, offering them a forum to present and discuss recent issues in methodological development as well as public health applications. It is divided into three parts, with the first providing an overview of Monte-Carlo techniques, the second focusing on missing data Monte-Carlo methods, and the third addressing Bayesian and general statistical modeling using Monte-Carlo simulations. The data and computer programs used here will also be made publicly available, allowing readers to replicate the model development and data analysis presented in each chapter, and to readily apply them in their own research. Featuring highly topical content, the book has the potential to impact model development and data analyses across a wide spectrum of fields, and to spark further research in this direction.

Chen, J., Tam, D., Raffel, C., Bansal, M., and Yang, D. (2021a). “An Empirical Survey of Data Augmentation for Limited Data Learning in NLP.” In: *arXiv e-Print*.

NLP has achieved great progress in the past decade through the use of neural models and large labeled datasets. The dependence on abundant data prevents NLP models from being applied to low-resource settings or novel tasks where significant time, money, or expertise is required to label massive amounts of textual data. Recently, data augmentation methods have been explored as a means of improving data efficiency in NLP. To date, there has been no systematic empirical overview of data augmentation for NLP in the limited labeled data setting, making it difficult to understand which methods work in which settings. In this paper, we provide an empirical survey of recent progress on data augmentation for NLP in the limited labeled data setting, summarizing the landscape of methods (including token-level augmentations, sentence-level augmentations, adversarial augmentations, and hidden-space augmentations) and carrying out experiments on 11 datasets covering topics/news classification, inference tasks, paraphrasing tasks, and single-sentence tasks. Based on the results, we draw several conclusions to help practitioners choose appropriate augmentations in different settings and discuss the current challenges and future directions for limited data learning in NLP.

Chen, W., Koo, B., Wang, Y., O’Hare, C., Langrené, N., Toscas, P., and Zhu, Z. (2021b). “Using a stochastic economic scenario generator to analyse uncertain superannuation and retirement outcomes.” In: *Annals of Actuarial Science*.

The retirement systems in many developed countries have been increasingly moving from defined benefit towards defined contribution system. In defined contribution systems, financial and longevity risks are shifted from pension providers to retirees. In this paper, we use a probabilistic approach to analyse the uncertainty associated with superannuation accumulation and decumulation. We apply an economic scenario generator called the Simulation of Uncertainty for Pension Analysis (SUPA) model to project uncertain future financial and economic variables. This multi-factor stochastic investment model, based on the Monte Carlo method, allows us to obtain the probability distribution of possible outcomes regarding the superannuation accumulation and decumulation phases, such as relevant percentiles. We present two examples to demonstrate the implementation of the SUPA model for the uncertainties during both phases under the current superannuation and Age Pension policy, and test two superannuation policy reforms suggested by the Grattan Institute.

Chen, W., Minney, A., Toscas, P., Koo, B., Zhu, Z., and Pantelous, A. A. (2021c). “Personalised drawdown strategies and partial annuitisation to mitigate longevity risk.” In: *Finance Research Letters* 39 (101644).

Despite the importance of drawdown strategies under a defined contribution system with increased longevity risk, little guidance to retired and retiring members has been forthcoming from superannuation funds. This paper provides a do-it-yourself drawdown design for members of superannuation funds along with comparison studies on a range of retirement income strategies under an array of realistic scenarios. A stochastic economic scenario generator is used to simulate the uncertain outcomes of different drawdown strategies during retirement. The impact of annuitisation for mitigating longevity risk under government pension rules and the selection of personalised drawdown and annuitisation strategies for retirement are examined.

Cheng, P.-K. and Planchet, F. (2018). “Stochastic Deflator for an Economic Scenario Generator with Five Factors.” In: *arXiv e-Print*.

In this paper, we implement a stochastic deflator with five economic and financial risk factors: interest rates, market price of risk, stock prices, default intensities, and convenience yields. We examine the deflator with different financial assets, such as stocks, zero-coupon bonds, vanilla options, and corporate coupon bonds. We find required regularity conditions to implement our stochastic deflator. Our numerical results show the reliability of the deflator approach in pricing financial derivatives.

Chevalier, C., Martius, O., and Ginsbourger, D. (2021). “Modeling Nonstationary Extreme Dependence With Sta-



tionary Max-Stable Processes and Multidimensional Scaling.” In: *Journal of Computational and Graphical Statistics*.

Modeling the joint distribution of extreme events at multiple locations is a challenging task with important applications. In this study, we use max-stable models to study extreme daily precipitation events in Switzerland. The nonstationarity of the spatial process at hand involves important challenges, which are often dealt with by using a stationary model in a so-called climate space, with well-chosen covariates. Here, we instead choose to warp the weather stations under study in a latent space of higher dimension using multidimensional scaling (MDS). Two methods are proposed to define target dissimilarity matrices, based respectively on extremal coefficients and on pairwise likelihoods. Results suggest that the proposed methods allow capturing complex spatial dependences of spatial extreme precipitations, enabling in turn to reliably extrapolate functionals such as extremal coefficients. Supplemental materials for this article are available online.

Chihara, L. M. and Hesterberg, T. C. (2018). *Mathematical Statistics with Resampling and R, 2nd Edition*. Wiley. 560 pp.

Resampling helps students understand the meaning of sampling distributions, sampling variability, P-values, hypothesis tests, and confidence intervals. The second edition of *Mathematical Statistics with Resampling and R* combines modern resampling techniques and mathematical statistics. This book has been classroom-tested to ensure an accessible presentation, uses the powerful and flexible computer language R for data analysis and explores the benefits of modern resampling techniques. This book offers an introduction to permutation tests and bootstrap methods that can serve to motivate classical inference methods. The book strikes a balance between theory, computing, and applications, and the new edition explores additional topics including consulting, paired t test, ANOVA and Google Interview Questions. Throughout the book, new and updated case studies are included representing a diverse range of subjects such as flight delays, birth weights of babies, and telephone company repair times. These illustrate the relevance of the real-world applications of the material. This new edition:

- Puts the focus on statistical consulting that emphasizes giving a client an understanding of data and goes beyond typical expectations
- Presents new material on topics such as the paired t test, Fisher’s Exact Test and the EM algorithm
- Offers a new section on “Google Interview Questions” that illustrates statistical thinking
- Provides a new chapter on ANOVA
- Contains more exercises and updated case studies, data sets, and R code

Written for undergraduate students in a mathematical statistics course as well as practitioners and researchers, the second edition of *Mathematical Statistics with Resampling and R* presents a revised and updated guide for applying the most current resampling techniques to mathematical statistics.

Clark, A. G., Walkinshaw, N., and Hierons, R. M. (2021). “Test case generation for agent-based models: A systematic literature review.” In: *arXiv e-Print*.

Agent-based models play an important role in simulating complex emergent phenomena and supporting critical decisions. In this context, a software fault may result in poorly informed decisions that lead to disastrous consequences. The ability to rigorously test these models is therefore essential. In this systematic literature review, we answer five research questions related to the key aspects of test case generation in agent-based models: What are the information artifacts used to generate tests? How are these tests generated? How is a verdict assigned to a generated test? How is the adequacy of a generated test suite measured? What level of abstraction of an agent-based model is targeted by a generated test? Our results show that whilst the majority of techniques are effective for testing functional requirements at the agent and integration levels of abstraction, there are comparatively few techniques capable of testing society-level behaviour. Additionally, we identify a need for more thorough evaluation using realistic case studies that feature challenging properties associated with a typical agent-based model.

Cogneau, P. and Zakamouline, V. (2013). “Block bootstrap methods and the choice of stocks for the long run.” In: *Quantitative Finance* 13(9), pp. 1443–1457.

Financial advisors commonly recommend that the investment horizon should be rather long in order to benefit from the “time diversification”. In this case, in order to choose the optimal portfolio, it is necessary to estimate the risk and reward of several alternative portfolios over a long-run given a sample of observations over a short-run. Two interrelated obstacles in these estimations are lack of sufficient data and the uncertainty in the nature of the return generating process. To overcome these obstacles researchers rely heavily on block bootstrap methods. In this paper we demonstrate that the estimates provided by a block bootstrap method are generally biased and we propose two methods of bias reduction. We show that an improper use of a block bootstrap method usually causes underestimation of the risk of a portfolio whose returns are independent over time and overestimation of the risk of a portfolio whose returns are mean-reverting.

- Cohort, P., Corbetta, J., and Laachir, I. (2020). “Analytical scores for stress scenarios.” In: *arXiv e-Print*.  
 In this work, inspired by the Archer-Mouy-Selmi approach, we present two methodologies for scoring the stress test scenarios used by CCPs for sizing their Default Funds. These methodologies can be used by risk managers to compare different sets of scenarios and could be particularly useful when evaluating the relevance of adding new scenarios to a pre-existing set.
- Coletta, A., Prata, M., Conti, M., Mercanti, E., Bartolini, N., Moulin, A., Vyetrenko, S., and Balch, T. (2021). “Towards Realistic Market Simulations: a Generative Adversarial Networks Approach.” In: *arXiv e-Print*.  
 Simulated environments are increasingly used by trading firms and investment banks to evaluate trading strategies before approaching real markets. Backtesting, a widely used approach, consists of simulating experimental strategies while replaying historical market scenarios. Unfortunately, this approach does not capture the market response to the experimental agents’ actions. In contrast, multi-agent simulation presents a natural bottom-up approach to emulating agent interaction in financial markets. It allows to set up pools of traders with diverse strategies to mimic the financial market trader population, and test the performance of new experimental strategies. Since individual agent-level historical data is typically proprietary and not available for public use, it is difficult to calibrate multiple market agents to obtain the realism required for testing trading strategies. To address this challenge we propose a synthetic market generator based on Conditional Generative Adversarial Networks (CGANs) trained on real aggregate-level historical data. A CGAN-based “world” agent can generate meaningful orders in response to an experimental agent. We integrate our synthetic market generator into ABIDES, an open source simulator of financial markets. By means of extensive simulations we show that our proposal outperforms previous work in terms of stylized facts reflecting market responsiveness and realism.
- Colson, A. R. and Cooke, R. M. (2017). “Cross validation for the classical model of structured expert judgment.” In: *Reliability Engineering and System Safety* 163, pp. 109–120.  
 Mathematical aggregation schemes are evaluated against data. Performance weighting is cross validated. Performance weighting is significantly better than equal weighting. We update the 2008 TU Delft structured expert judgment database with data from 33 professionally contracted Classical Model studies conducted between 2006 and March 2015 to evaluate its performance relative to other expert aggregation models. We briefly review alternative mathematical aggregation schemes, including harmonic weighting, before focusing on linear pooling of expert judgments with equal weights and performance-based weights. Performance weighting outperforms equal weighting in all but 1 of the 33 studies in-sample. True out-of-sample validation is rarely possible for Classical Model studies, and cross validation techniques that split calibration questions into a training and test set are used instead. Performance weighting incurs an “out-of-sample penalty” and its statistical accuracy out-of-sample is lower than that of equal weighting. However, as a function of training set size, the statistical accuracy of performance-based combinations reaches 75 percent of the equal weight value when the training set includes 80 percent of calibration variables. At this point the training set is sufficiently powerful to resolve differences in individual expert performance. The information of performance-based combinations is double that of equal weighting when the training set is at least 50 percent of the set of calibration variables. Previous out-of-sample validation work used a Total Out-of-Sample Validity Index based on all splits of the calibration questions into training and test subsets, which is expensive to compute and includes small training sets of dubious value. As an alternative, we propose an Out-of-Sample Validity Index based on averaging the product of statistical accuracy and information over all training sets sized at 80 percent of the calibration set. Performance weighting outperforms equal weighting on this Out-of-Sample Validity Index in 26 of the 33 post-2006 studies; the probability of 26 or more successes on 33 trials if there were no difference between performance weighting and equal weighting is 0.001.
- Cramer, E., Gorjao, L. R., Mitsos, A., Schafer, B., Witthaut, D., and Dahmen, M. (2021). “Validation Methods for Energy Time Series Scenarios from Deep Generative Models.” In: *arXiv e-Print*.  
 The design and operation of modern energy systems are heavily influenced by time-dependent and uncertain parameters, e.g., renewable electricity generation, load-demand, and electricity prices. These are typically represented by a set of discrete realizations known as scenarios. A popular scenario generation approach uses deep generative models (DGM) that allow scenario generation without prior assumptions about the data distribution. However, the validation of generated scenarios is difficult, and a comprehensive discussion about appropriate validation methods is currently lacking. To start this discussion, we provide a critical assessment of the currently used validation methods in the energy scenario generation literature. In particular, we assess validation methods based on probability density, auto-correlation, and power spectral density. Furthermore, we propose using the multifractal detrended fluctuation analysis (MFDFA) as an additional validation method for non-trivial features



like peaks, bursts, and plateaus. As representative examples, we train generative adversarial networks (GANs), Wasserstein GANs (WGANs), and variational autoencoders (VAEs) on two renewable power generation time series (photovoltaic and wind from Germany in 2013 to 2015) and an intra-day electricity price time series from the European Energy Exchange in 2017 to 2019. We apply the four validation methods to both the historical and the generated data and discuss the interpretation of validation results as well as common mistakes, pitfalls, and limitations of the validation methods. Our assessment shows that no single method sufficiently characterizes a scenario but ideally validation should include multiple methods and be interpreted carefully in the context of scenarios over short time periods.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. (2019). “[RandAugment: Practical data augmentation with no separate search.](#)” In: *arXiv e-Print*.

Recent work has shown that data augmentation has the potential to significantly improve the generalization of deep learning models. Recently, learned augmentation strategies have led to state-of-the-art results in image classification and object detection. While these strategies were optimized for improving validation accuracy, they also led to state-of-the-art results in semi-supervised learning and improved robustness to common corruptions of images. One obstacle to a large-scale adoption of these methods is a separate search phase which significantly increases the training complexity and may substantially increase the computational cost. Additionally, due to the separate search phase, these learned augmentation approaches are unable to adjust the regularization strength based on model or dataset size. Learned augmentation policies are often found by training small models on small datasets and subsequently applied to train larger models. In this work, we remove both of these obstacles. RandAugment may be trained on the model and dataset of interest with no need for a separate proxy task. Furthermore, due to the parameterization, the regularization strength may be tailored to different model and dataset sizes. RandAugment can be used uniformly across different tasks and datasets and works out of the box, matching or surpassing all previous learned augmentation approaches on CIFAR-10, CIFAR-100, SVHN, and ImageNet. On the ImageNet dataset we achieve 85.0% accuracy, a 0.6% increase over the previous state-of-the-art and 1.0% increase over baseline augmentation. On object detection, RandAugment leads to 1.0-1.3% improvement over baseline augmentation, and is within 0.3% mAP of AutoAugment on COCO. Finally, due to its interpretable hyperparameter, RandAugment may be used to investigate the role of data augmentation with varying model and dataset size.

Czado, C. and Nagler, T. (2022). “[Vine Copula Based Modeling.](#)” In: *Annual Review of Statistics and Its Application* 9(1).

With the availability of massive multivariate data comes a need to develop flexible multivariate distribution classes. The copula approach allows marginal models to be constructed for each variable separately and joined with a dependence structure characterized by a copula. The class of multivariate copulas was limited for a long time to elliptical (including the Gaussian and t-copula) and Archimedean families (such as Clayton and Gumbel copulas). Both classes are rather restrictive with regard to symmetry and tail dependence properties. The class of vine copulas overcomes these limitations by building a multivariate model using only bivariate building blocks. This gives rise to highly flexible models that still allow for computationally tractable estimation and model selection procedures. These features made vine copula models quite popular among applied researchers in numerous areas of science. This article reviews the basic ideas underlying these models, presents estimation and model selection approaches, and discusses current developments and future directions.

Czaronis, M., Kritzman, M., Pamir, B., and Turkington, D. (2020). “[Enhanced Scenario Analysis.](#)” In: *The Journal of Portfolio Management* 46(4), pp. 69–79.

Investors have long relied on scenario analysis as an alternative to mean-variance analysis to help them construct portfolios. Even though mean-variance analysis accounts for all potential scenarios, many investors find it difficult to implement because it requires them to specify statistical features of asset classes that are often unintuitive and difficult to estimate. Scenario analysis, by contrast, requires only that investors specify a small set of potential outcomes as projections of economic variables and assign probabilities to their occurrence. It is, therefore, more intuitive than mean-variance analysis, but it is highly subjective. In this article, the authors propose to replace the subjective elements of scenario analysis with a robust statistical process. They use a multivariate measure of statistical distance to estimate probabilities of prospective scenarios. Next, they construct portfolios that maximize utility for investors with different risk preferences. Last, the authors introduce a procedure for minimally modifying scenarios to render them consistent with prespecified views about their probabilities of occurrence.

Czaronis, M., Kritzman, M., and Turkington, D. (2021). “[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.

- Da Silva, B. and Shi, S. S. (2019). “Towards Improved Generalization in Financial Markets with Synthetic Data Generation.” In: *arXiv e-Print*.

Training deep learning models that generalize well to live deployment is a challenging problem in the financial markets. The challenge arises because of high dimensionality, limited observations, changing data distributions, and a low signal-to-noise ratio. High dimensionality can be dealt with using robust feature selection or dimensionality reduction, but limited observations often result in a model that overfits due to the large parameter space of most deep neural networks. We propose a generative model for financial time series, which allows us to train deep learning models on millions of simulated paths. We show that our generative model is able to create realistic paths that embed the underlying structure of the markets in a way stochastic processes cannot.

- Dahl, C. M. and Sorensen, E. N. (2021). “Time Series (re)sampling using Generative Adversarial Networks.” In: *arXiv e-Print*.

We propose a novel bootstrap procedure for dependent data based on Generative Adversarial networks (GANs). We show that the dynamics of common stationary time series processes can be learned by GANs and demonstrate that GANs trained on a single sample path can be used to generate additional samples from the process. We find that temporal convolutional neural networks provide a suitable design for the generator and discriminator, and that convincing samples can be generated on the basis of a vector of iid normal noise. We demonstrate the finite sample properties of GAN sampling and the suggested bootstrap using simulations where we compare the performance to circular block bootstrapping in the case of resampling an AR(1) time series processes. We find that resampling using the GAN can outperform circular block bootstrapping in terms of empirical coverage.

- Davidson, R. (2017). “Diagnostics for the bootstrap and fast double bootstrap.” In: *Econometric Reviews* 36(6-9), pp. 1021–1038.

The bootstrap is typically much less reliable in the context of time-series models with serial correlation of unknown form than it is when regularity conditions for the conventional IID bootstrap, based on resampling, apply. It is therefore useful for practitioners to have available diagnostic techniques capable of evaluating bootstrap performance.

- Davis, M. and Lleo, S. (2015). “Behaviouralizing Black-Litterman: Expert Opinions and Behavioural Biases in a Diffusion Setting.” In: *SSRN e-Print*.

This paper proposes a continuous time version of the Black-Litterman model that accounts for, and corrects, some of the main behavioural biases that analysts may exhibit. We formulate the model as a stochastic control problem under partial observations, and derive the optimal investment strategy and value function in closed form. We implement this model with three partially observable state variables corresponding to the three factors of the Fama-French model, and fourteen sources of observations: market data from eleven ETFs, views from two analysts, and one stress test scenario. With this example, we show concretely how to calibrate analyst views and mitigate the impact of behavioural biases. To explore the effect that views and biases have on the asset allocation, we compare the results of six dynamic investment models. We find that the views have a modest impact on the Kelly portfolio, but the confidence intervals around the views have a large impact on the intertemporal hedging portfolio.

- Davis, M. H. A. and Lleo, S. (2016). “A Simple Procedure for Combining Expert Opinion with Statistical Estimates to Achieve Superior Portfolio Performance.” In: *The Journal of Portfolio Management* 42(4), pp. 49–58.

In this article, the authors describe a simple procedure for combining statistical estimates with expert opinions to produce a view of future asset performance. The authors discuss the impact of behavioral bias on these views and propose general modeling principles to reduce this bias. They use standard linear filtering techniques to combine statistical estimates with expert opinions seamlessly and discuss applications to dynamic portfolio optimization.

- De Gennaro Aquino, L., Sornette, D., and Strub, M. S. (2021). “Portfolio Selection With Exploration of New Investment Opportunities.” In: *SSRN e-Print*.

We introduce a model for portfolio selection with an extendable investment universe where an agent with mean-variance preferences faces a trade-off between exploiting existing and exploring for new investment opportunities. When the agent chooses to explore, a new risky asset is discovered and the agent subsequently invests in the extended universe. We first provide conditions for wellposedness and characterize the optimal amount devoted to exploration. Our model shows that incremental exploration does not pay off, that it is increasingly worthwhile

to explore in worse market environments, and that investment performance measured by the Sharpe ratio is increasing in the initial wealth of the agent indicating that richer agents can make better use of new investment opportunities. We further generalize our model and verify the robustness of the main findings with regards to various modeling assumptions.

- De Luca, G. and Zuccolotto, P. (2016). “A double clustering algorithm for financial time series based on extreme events.” In: *Statistics and Risk Modeling* 0(0).

This paper is concerned with a procedure for financial time series clustering, aimed at creating groups of time series characterized by similar behavior with regard to extreme events. The core of our proposal is a double clustering procedure: the former is based on the lower tail dependence of all the possible pairs of time series, the latter on the upper tail dependence. Tail dependence coefficients are estimated with copula functions. The final goal is to exploit the two clustering solutions in an algorithm designed to create a portfolio that maximizes the probability of joint positive extreme returns while minimizing the risk of joint negative extreme returns. In financial crisis scenarios, such a portfolio is expected to outperform portfolios generated by the traditional methods. We describe the results of a simulation study and, finally, we apply the procedure to a dataset composed of the 50 assets included in the EUROSTOXX index.

- De Meer, F., Schwendner, P., and Wunsch, M. (2021). “Tackling the exponential scaling of signature-based GANs for high-dimensional financial time series generation.” In: *SSRN e-Print*.

Generative Adversarial Networks (GANs) have been shown to be able to generate samples of complex financial time series, particularly by employing the concept of path signatures, a universal description of the geometric properties of a data stream whose expected value uniquely characterizes the time series.

In particular, the SigCWGAN model is able to generate time series of arbitrary length; however, the parameters of the neural network employed grow exponentially with the dimension of the underlying time series, which makes the model intractable when seeking to generate large financial market scenarios.

To overcome this curse of dimensionality, we propose an iterative generation procedure relying on the concept of hierarchies in financial markets. We construct an ensemble of GANs, that we call the Hierarchical-SigCWGAN, based on hierarchical clustering that approximate signatures in the spirit of the original model.

The Hierarchical-SigCWGAN is able to scale to higher dimensions and generate large-dimensional scenarios in which the joint behavior of all the assets in the market is replicated. We validate our model by comparing its performance on a series of similarity metrics with respect to the original SigCWGAN on a dataset where it is still tractable and showcase its scalability.

- De Meer Pardo, F. (2019). “Enriching Financial Datasets with Generative Adversarial Networks.” MA thesis. Delft University of Technology.

The scarcity of historical financial data has been a huge hindrance for the development algorithmic trading models ever since the first models were devised. Most financial models assume as hypothesis a series of characteristics regarding the nature of financial time series and seek extracting information about the state of the market through calibration. Through backtesting, a large number of these models are seen not to perform and are thus discarded. The remaining well-performing models however, are highly vulnerable to overfitting. Financial time series are complex by nature and their behaviour changes over time, so this concern is well founded. In addition to the problem of overfitting, available data is far too scarce for most machine learning applications and impossibly scarce for advanced approaches such as reinforcement learning, which has heavily impaired the application of these novel techniques in financial settings. This is where data generation comes into play. Generative Adversarial Networks, GANs, are a type of neural network architecture that focuses on sample generation. Through adversarial training, the GAN can learn the underlying structure of the input data and become able to generate samples very similar to those of the data distribution. This is specially useful in the case of high-dimensional objects, in which the dimensions are heavily inter-dependent, such as images, music and in our case financial time series. In this work we want to explore the generating capabilities of GANs applied to financial time series and investigate whether or not we can generate realistic financial scenarios.

- De Meer Pardo, F. and Lopez, R. C. (2020). “Mitigating Overfitting on Financial Datasets with Generative Adversarial Networks.” In: *The Journal of Financial Data Science* 2 (1), pp. 76–85.

Overfitting is an inevitable phenomenon when applying deep learning techniques to financial data, given the relative scarcity of available historical data and the ever-changing nature of financial series. This seemingly unavoidable pitfall has heavily impaired the application of many machine learning techniques, such as deep or reinforcement learning, to financial settings. Data augmentation can be an option to circumvent this, specifically generative adversarial networks (GANs). GANs are a type of neural network architecture that focuses on sample

generation. Through adversarial training, the GAN can implicitly learn the underlying structure inherent to the dynamics of financial series and acquire the capacity to generate scenarios that share many similarities to those seen in the historic time series. In this article, the authors propose a data augmentation technique using the Wasserstein GAN with gradient penalty and show how training deep models on synthetic data mitigates overfitting, improving their performance on test data when compared to models trained solely on real data.

De Meo, E. (2019). “[Scenario Design for Macro-Financial Stress Testing](#).” In: *SSRN e-Print*.

The goal of this paper is to provide a possible approach to Scenario Design for selecting a stress scenario on economic growth, inflation and long-term interest rates in Italy. The Scenario Design framework belongs to the class of Second Generation Stress Tests and is composed of a few building blocks. First, multiple scenarios on the risk factors are generated simulating a Large Bayesian VAR for the Italian economy. Second, we take the perspective of a representative investor who aims to select a severe yet plausible scenario on the systematic risk factors following a factor investing strategy. Moreover, we compare the stress scenarios selected under two different approaches to measure plausibility: the Mahalanobis distance and Entropy pooling under three alternative subjective views with a clear economic narrative. We give evidence that our framework is suitable for the selection of a proper forward-looking severe yet plausible stress scenario.

de Miranda Cardoso, J. V., Ying, J., and Palomar, D. P. (2020). “[Algorithms for Learning Graphs in Financial Markets](#).” In: *arXiv e-Print*.

In the past two decades, the field of applied finance has tremendously benefited from graph theory. As a result, novel methods ranging from asset network estimation to hierarchical asset selection and portfolio allocation are now part of practitioners’ toolboxes. In this paper, we investigate the fundamental problem of learning undirected graphical models under Laplacian structural constraints from the point of view of financial market times series data. In particular, we present natural justifications, supported by empirical evidence, for the usage of the Laplacian matrix as a model for the precision matrix of financial assets, while also establishing a direct link that reveals how Laplacian constraints are coupled to meaningful physical interpretations related to the market index factor and to conditional correlations between stocks. Those interpretations lead to a set of guidelines that practitioners should be aware of when estimating graphs in financial markets. In addition, we design numerical algorithms based on the alternating direction method of multipliers to learn undirected, weighted graphs that take into account stylized facts that are intrinsic to financial data such as heavy tails and modularity. We illustrate how to leverage the learned graphs into practical scenarios such as stock time series clustering and foreign exchange network estimation. The proposed graph learning algorithms outperform the state-of-the-art methods in an extensive set of practical experiments. Furthermore, we obtain theoretical and empirical convergence results for the proposed algorithms. Along with the developed methodologies for graph learning in financial markets, we release an R package, called *fingraph*, accommodating the code and data to obtain all the experimental 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 off-the-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.

Denev, A. and Mutnikas, Y. (2016). “[A Formalized, Integrated and Visual Approach to Stress Testing](#).” In: *SSRN e-Print*.

In this paper we will give for the first time a formal mathematical language to the steps used currently by financial institutions when calculating the impact of a stress scenario on a balance sheet that depends on more granular or different factors than those provided in the scenario. We will introduce the language of Probabilistic Graphical Models (PGM) a technique rooted in machine learning to show how the different models used at each step can be put together in a coherent picture thus giving a holistic view of the entire models setup. This will give us a solid basis to discuss some weaknesses and problems with the stress testing exercises run by the industry as of today. We will show empirical analyses to substantiate better some of our claims.

- Diffenbaugh, N. S. (2020). “[Verification of extreme event attribution: Using out-of-sample observations to assess changes in probabilities of unprecedented events.](#)” In: *Science Advances* 6(12), eaay2368.

Independent verification of anthropogenic influence on specific extreme climate events remains elusive. This study presents a framework for such verification. This framework reveals that previously published results based on a 1961-2005 attribution period frequently underestimate the influence of global warming on the probability of unprecedented extremes during the 2006-2017 period. This underestimation is particularly pronounced for hot and wet events, with greater uncertainty for dry events. The underestimation is reflected in discrepancies between probabilities predicted during the attribution period and frequencies observed during the out-of-sample verification period. These discrepancies are most explained by increases in climate forcing between the attribution and verification periods, suggesting that 21st-century global warming has substantially increased the probability of unprecedented hot and wet events. Hence, the use of temporally lagged periods for attribution-and, more broadly, for extreme event probability quantification-can cause underestimation of historical impacts, and current and future risks.

- Ding, D., Zhang, M., Pan, X., Yang, M., and He, X. (2019). “[Modeling Extreme Events in Time Series Prediction.](#)” In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM.

Time series prediction is an intensively studied topic in data mining. In spite of the considerable improvements, recent deep learning-based methods overlook the existence of extreme events, which result in weak performance when applying them to real time series. Extreme events are rare and random, but do play a critical role in many real applications, such as the forecasting of financial crisis and natural disasters. In this paper, we explore the central theme of improving the ability of deep learning on modeling extreme events for time series prediction. Through the lens of formal analysis, we first find that the weakness of deep learning methods roots in the conventional form of quadratic loss. To address this issue, we take inspirations from the Extreme Value Theory, developing a new form of loss called Extreme Value Loss (EVL) for detecting the future occurrence of extreme events. Furthermore, we propose to employ Memory Network in order to memorize extreme events in historical records. By incorporating EVL with an adapted memory network module, we achieve an end-to-end framework for time series prediction with extreme events. Through extensive experiments on synthetic data and two real datasets of stock and climate, we empirically validate the effectiveness of our framework. Besides, we also provide a proper choice for hyper-parameters in our proposed framework by conducting several additional experiments.

- Dogariu, M., Stefan, L.-D., Boteanu, B. A., Lamba, C., and Bogdan Ionescu, B. K. nd (2021a). “[Generation of Realistic Synthetic Financial Time-Series.](#)” In: *ACM Trans. Multimedia Comput. Commun. Appl* 37(4) (111).

Financial markets have always been a point of interest for automated systems. Due to their complex nature, financial algorithms and fintech frameworks require vast amounts of data to accurately respond to market fluctuations. This data availability is tied to the daily market evolution so it is impossible to accelerate its acquisition. In this paper, we discuss several solutions for augmenting financial datasets via synthesizing realistic time-series with the help of generative models. This problem is complex since financial time series present very specific properties, e.g., fat-tail distribution, cross-correlation between different stocks, specific autocorrelation, cluster volatility etc. In particular, we propose solutions for capturing cross-correlations between different stocks and for transitioning from fixed to variable length time-series without resorting to sequence modeling networks, and adapt various network architectures, e.g., fully connected and convolutional GANs, variational autoencoders, and generative moment matching networks. Finally, we tackle the problem of evaluating the quality of synthetic financial time-series. We introduce qualitative and quantitative metrics, along with a portfolio trend prediction framework which validates our generative models’ performance. We carry out experiments on real-world financial data extracted from the US stock market proving the benefits of these techniques.

- Dogariu, M., Stefan, L.-D., Boteanu, B. A., Lamba, C., and Ionescu, B. (2021b). “[Towards Realistic Financial Time Series Generation via Generative Adversarial Learning.](#)” In: *European Signal Processing Conference - EUSIPCO*.



Training network models to accurately respond to market fluctuations requires access to vast amounts of data. Data availability is strictly bound to the market’s evolution, which updates only on a daily basis. In this paper, we propose several solutions based on Generative Adversarial Networks for providing artificially generated time series data with realistic properties. The main challenge here is the specificity of the target data, which has properties that are difficult to control and have wide variations in time, e.g., central moment statistics, autocorrelation or cluster volatility. Another contribution is in assessing the quality of synthetic data, in general, as there is no standard metric for this. Experimental validation is carried out on real-world financial data retrieved from the US stock market.

Eckerli, F. (2021). “Generative Adversarial Networks in finance: an overview.” In: *SSRN e-Print*.

Modelling in finance is a challenging task: the data often has complex statistical properties and its inner workings are largely unknown. Deep learning algorithms are making progress in the field of data-driven modelling, but the lack of sufficient data to train these models is currently holding back several new applications. Generative Adversarial Networks (GANs) are a neural network architecture family that has achieved good results in image generation and is being successfully applied to generate time series and other types of financial data. The purpose of this study is to present an overview of how these GANs work, their capabilities and limitations in the current state of research with financial data, and present some practical applications in the industry. As a proof of concept, three known GAN architectures were tested on financial time series, and the generated data was evaluated on its statistical properties, yielding solid results. Finally, it was shown that GANs have made considerable progress in their finance applications and can be a solid additional tool for data scientists in this field.

Eckerli, F. and Osterrieder, J. (2021). “Generative Adversarial Networks in finance: an overview.” In: *arXiv e-Print*.

Modelling in finance is a challenging task: the data often has complex statistical properties and its inner workings are largely unknown. Deep learning algorithms are making progress in the field of data-driven modelling, but the lack of sufficient data to train these models is currently holding back several new applications. Generative Adversarial Networks (GANs) are a neural network architecture family that has achieved good results in image generation and is being successfully applied to generate time series and other types of financial data. The purpose of this study is to present an overview of how these GANs work, their capabilities and limitations in the current state of research with financial data, and present some practical applications in the industry. As a proof of concept, three known GAN architectures were tested on financial time series, and the generated data was evaluated on its statistical properties, yielding solid results. Finally, it was shown that GANs have made considerable progress in their finance applications and can be a solid additional tool for data scientists in this field.

El Karoui, N. and Purdom, E. (2018). “Can we trust the bootstrap in high-dimensions? the case of linear models.” In: *The Journal of Machine Learning Research*.

We consider the performance of the bootstrap in high-dimensions for the setting of linear regression, where  $p < n$  but  $p/np/n$  is not close to zero. We consider ordinary least-squares as well as robust regression methods and adopt a minimalist performance requirement: can the bootstrap give us good confidence intervals for a single coordinate of beta (where beta is the true regression vector)? We show through a mix of numerical and theoretical work that the bootstrap is fraught with problems. Both of the most commonly used methods of bootstrapping for regression-residual bootstrap and pairs bootstrap-give very poor inference on beta as the ratio  $p/np/n$  grows. We find that the residual bootstrap tend to give anti-conservative estimates (inflated Type I error), while the pairs bootstrap gives very conservative estimates (severe loss of power) as the ratio  $p/np/n$  grows. We also show that the jackknife resampling technique for estimating the variance of  $\beta_{\text{hat}}$  severely overestimates the variance in high dimensions. We contribute alternative procedures based on our theoretical results that result in dimensionality adaptive and robust bootstrap methods.

Elkamhi, R. and Stefanova, D. (2015). “Dynamic Hedging and Extreme Asset Co-movements.” In: *The Review of Financial Studies* 28(3), pp. 743–790.

The paper investigates the portfolio allocation effects of increased asset co-movements during market downturns. We develop a model for the stock price process that allows for increased and asymmetric dependence between extreme return realizations. We isolate the portfolio hedging demands that arise due to extreme co-movements and find a substantial shift of the portfolio holdings toward the risk-free asset. We demonstrate that accounting for dependence between extreme events in portfolio decisions leads to significant economic gains that stem primarily from intertemporal hedging motives. These findings are robust along alternative modeling assumptions of extreme co-movements and conditional correlation.

Engelke, S. and Ivanovs, J. (2021). “Sparse Structures for Multivariate Extremes.” In: *Annual Review of Statistics and Its Application* 8(1).

Extreme value statistics provides accurate estimates for the small occurrence probabilities of rare events. While theory and statistical tools for univariate extremes are well developed, methods for high-dimensional and complex data sets are still scarce. Appropriate notions of sparsity and connections to other fields such as machine learning, graphical models, and high-dimensional statistics have only recently been established. This article reviews the new domain of research concerned with the detection and modeling of sparse patterns in rare events. We first describe the different forms of extremal dependence that can arise between the largest observations of a multivariate random vector. We then discuss the current research topics, including clustering, principal component analysis, and graphical modeling for extremes. Identification of groups of variables that can be concomitantly extreme is also addressed. The methods are illustrated with an application to flood risk assessment.

Engle, R. F. (2020). “Stress Testing with Market Data.” In: *SSRN e-Print*.

A stress test assesses the value of a firm or asset in the future under an adverse counterfactual scenario. The critical points of stress tests are the valuation model and the scenario. This paper describes some of the difficulties in generating appropriate scenarios and valuing firms under these scenarios. In most cases, these difficulties can be solved if the regulator is better informed than the market. However, if this is not correct at all times and settings, then it is also sensible to carry out stress tests with market scenarios and market data. When these stress tests agree, the results gain added credibility. When they disagree, the parties can discuss whether the market has missed signals, or whether the regulators’ models are wrong or have been politically impacted. Detailed analysis of SRISK, a market based stress test, is presented from an economic, econometric and historical point of view. This is compared with alternative measures such as SES and CoVaR and with regulatory stress tests.

Esmaeili, A., Gallagher, J. C., Springer, J. A., and Matson, E. T. (2021). “HAMLET: A Hierarchical Agent-based Machine Learning Platform.” In: *ACM Transactions on Autonomous and Adaptive Systems* 16(3-4), pp. 1–46.

Hierarchical Multi-agent Systems provide convenient and relevant ways to analyze, model, and simulate complex systems composed of a large number of entities that interact at different levels of abstraction. In this article, we introduce HAMLET (Hierarchical Agent-based Machine LEarning plaTform), a hybrid machine learning platform based on hierarchical multi-agent systems, to facilitate the research and democratization of geographically and/or locally distributed machine learning entities. The proposed system models machine learning solutions as a hypergraph and autonomously sets up a multi-level structure of heterogeneous agents based on their innate capabilities and learned skills. HAMLET aids the design and management of machine learning systems and provides analytical capabilities for research communities to assess the existing and/or new algorithms/datasets through flexible and customizable queries. The proposed hybrid machine learning platform does not assume restrictions on the type of learning algorithms/datasets and is theoretically proven to be sound and complete with polynomial computational requirements. Additionally, it is examined empirically on 120 training and 4 generalized batch testing tasks performed on 24 machine learning algorithms and 9 standard datasets. The provided experimental results not only establish confidence in the platform’s consistency and correctness but also demonstrate its testing and analytical capacity.

Facchinato, S. and Pola, G. (2014). “Managing uncertainty with diversification across macroeconomic scenarios (DAMS): from asset segmentation to portfolio.” In: *SSRN e-Print*.

Recent history has provided an excellent laboratory to test the robustness of investment processes. Despite claims of diversification, most balanced portfolios and pension funds were concentrated on equity risk and, consequently, key investment decisions ultimately consisted in a single binary bet: buy or sell equity. This led to pro-cyclical returns and generated a broad debate on the effectiveness of active management in generating performance in difficult market conditions. In 2011 AMUNDI Italy decided to revise the asset allocation process starting with a reinterpretation of portfolio diversification in terms of Diversification Across Macroeconomic Scenarios (DAMS). The main ambitions of DAMS are: (i) to explain complex patterns of large investment universes in terms of a limited number of factors and (ii) to catch up the market risk premium without being exposed to specific macroeconomic dynamics and asset idiosyncratic risk. In a previous study we illustrated the DAMS principle and implications in terms of asset segmentation. The aim of this paper is to move towards a new framework for multi-asset portfolio management, what we call DAMS second generation. DAMS first generation is enriched with new concepts and tools that enable us (i) to infer market expectations on relevant macroeconomic factors (growth and inflation) and global risk premium, and (ii) to properly manage portfolios via strategic and tactical asset allocation.



Faez, F., Ommi, Y., Baghshah, M. S., and Rabiee, H. R. (2020). “Deep Graph Generators: A Survey.” In: *arXiv e-Print*.

Deep generative models have achieved great success in areas such as image, speech, and natural language processing in the past few years. Thanks to the advances in graph-based deep learning, and in particular graph representation learning, deep graph generation methods have recently emerged with new applications ranging from discovering novel molecular structures to modeling social networks. This paper conducts a comprehensive survey on deep learning-based graph generation approaches and classifies them into five broad categories, namely, autoregressive, autoencoder-based, RL-based, adversarial, and flow-based graph generators, providing the readers a detailed description of the methods in each class. We also present publicly available source codes, commonly used datasets, and the most widely utilized evaluation metrics. Finally, we highlight the existing challenges and discuss future research directions.

Faggini, M., Bruno, B., and Parziale, A. (2019). “Crises in economic complex networks: Black Swans or Dragon Kings?” In: *Economic Analysis and Policy* 62, pp. 105–115.

Abstract After the global financial crisis of 2008, the literature paid new attention to economic crises, analysing them according to a network perspective. Assuming this perspective, this paper is aimed at offering an overview of the functioning of the global economy as a complex network, characterized by cascading failures when an extreme event (EE) occurs, and showing that the economic crisis of 2008 was an extreme event with well-identified features and consequently was forecastable.

Fakoor, R., Mueller, J., Erickson, N., Chaudhari, P., and Smola, A. J. (2020). “Fast, Accurate, and Simple Models for Tabular Data via Augmented Distillation.” In: *arXiv e-Print*.

Automated machine learning (AutoML) can produce complex model ensembles by stacking, bagging, and boosting many individual models like trees, deep networks, and nearest neighbor estimators. While highly accurate, the resulting predictors are large, slow, and opaque as compared to their constituents. To improve the deployment of AutoML on tabular data, we propose FAST-DAD to distill arbitrarily complex ensemble predictors into individual models like boosted trees, random forests, and deep networks. At the heart of our approach is a data augmentation strategy based on Gibbs sampling from a self-attention pseudolikelihood estimator. Across 30 datasets spanning regression and binary/multiclass classification tasks, FAST-DAD distillation produces significantly better individual models than one obtains through standard training on the original data. Our individual distilled models are over 10x faster and more accurate than ensemble predictors produced by AutoML tools like H2O/AutoSklearn.

Fang, J. and Lin, J. (2020). “Prior knowledge distillation based on financial time series.” In: *arXiv e-Print*.

One of the major characteristics of financial time series is that they contain a large amount of non-stationary noise, which is challenging for deep neural networks. People normally use various features to address this problem. However, the performance of these features depends on the choice of hyper-parameters. In this paper, we propose to use neural networks to represent these indicators and train a large network constructed of smaller networks as feature layers to fine-tune the prior knowledge represented by the indicators. During back propagation, prior knowledge is transferred from human logic to machine logic via gradient descent. Prior knowledge is the deep belief of neural network and teaches the network to not be affected by non-stationary noise. Moreover, co-distillation is applied to distill the structure into a much smaller size to reduce redundant features and the risk of overfitting. In addition, the decisions of the smaller networks in terms of gradient descent are more robust and cautious than those of large networks. In numerical experiments, we find that our algorithm is faster and more accurate than traditional methods on real financial datasets. We also conduct experiments to verify and comprehend the method.

Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., and Muller, P.-A. (2018). “Data augmentation using synthetic data for time series classification with deep residual networks.” In: *arXiv e-Print*.

Data augmentation in deep neural networks is the process of generating artificial data in order to reduce the variance of the classifier with the goal to reduce the number of errors. This idea has been shown to improve deep neural network’s generalization capabilities in many computer vision tasks such as image recognition and object localization. Apart from these applications, deep Convolutional Neural Networks (CNNs) have also recently gained popularity in the Time Series Classification (TSC) community. However, unlike in image recognition problems, data augmentation techniques have not yet been investigated thoroughly for the TSC task. This is surprising as the accuracy of deep learning models for TSC could potentially be improved, especially for small datasets that exhibit overfitting, when a data augmentation method is adopted. In this paper, we fill this gap by investigating the application of a recently proposed data augmentation technique based on the Dynamic Time

Warping distance, for a deep learning model for TSC. To evaluate the potential of augmenting the training set, we performed extensive experiments using the UCR TSC benchmark. Our preliminary experiments reveal that data augmentation can drastically increase deep CNN’s accuracy on some datasets and significantly improve the deep model’s accuracy when the method is used in an ensemble approach.

Flaig, S. and Junike, G. (2022). “[Scenario generation for market risk models using generative neural networks.](#)” In: *arXiv e-Print*.

In this research, we show how to expand existing approaches of generative adversarial networks (GANs) being used as economic scenario generators (ESG) to a whole internal model - with enough risk factors to model the full band-width of investments for an insurance company and for a one year horizon as required in Solvency 2. For validation of this approach as well as for optimisation of the GAN architecture, we develop new performance measures and provide a consistent, data-driven framework. Finally, we demonstrate that the results of a GAN-based ESG are similar to regulatory approved internal models in Europe. Therefore, GAN-based models can be seen as an assumption-free data-driven alternative way of market risk modelling.

Flood, M. D. and Korenko, G. G. (2015). “[Systematic scenario selection: stress testing and the nature of uncertainty.](#)” In: *Quantitative Finance* 15(1), pp. 43–59.

We present a technique for selecting multidimensional shock scenarios for use in financial stress testing. The methodology systematically enforces internal consistency among the shock dimensions by sampling points of arbitrary severity from a plausible joint probability distribution. The approach involves a grid search of sparse, well distributed, stress-test scenarios, which we regard as a middle ground between traditional stress testing and reverse stress testing. Choosing scenarios in this way reduces the danger of blind spots’ in stress testing. We suggest extensions to address the issues of non-monotonic loss functions and univariate shocks. We provide tested and commented source code in Matlab.

Fons, E., Dawson, P., Zeng, X.-j., Keane, J., and Iosifidis, A. (2021a). “[Adaptive Weighting Scheme for Automatic Time-Series Data Augmentation.](#)” In: *arXiv e-Print*.

Data augmentation methods have been shown to be a fundamental technique to improve generalization in tasks such as image, text and audio classification. Recently, automated augmentation methods have led to further improvements on image classification and object detection leading to state-of-the-art performances. Nevertheless, little work has been done on time-series data, an area that could greatly benefit from automated data augmentation given the usually limited size of the datasets. We present two sample-adaptive automatic weighting schemes for data augmentation: the first learns to weight the contribution of the augmented samples to the loss, and the second method selects a subset of transformations based on the ranking of the predicted training loss. We validate our proposed methods on a large, noisy financial dataset and on time-series datasets from the UCR archive. On the financial dataset, we show that the methods in combination with a trading strategy lead to improvements in annualized returns of over 50%, and on the time-series data we outperform state-of-the-art models on over half of the datasets, and achieve similar performance in accuracy on the others.

Fons, E., Dawson, P., Zeng, X.-j., Keane, J., and Iosifidis, A. (2021b). “[Augmenting Transferred Representations for Stock Classification.](#)” In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Stock classification is a challenging task due to high levels of noise and volatility of stocks returns. In this paper we show that using transfer learning can help with this task, by pre-training a model to extract universal features on the full universe of stocks of the S&P500 index and then transferring it to another model to directly learn a trading rule. Transferred models present more than double the risk-adjusted returns than their counterparts trained from zero. In addition, we propose the use of data augmentation on the feature space defined as the output of a pre-trained model (i.e. augmenting the aggregated time-series representation). We compare this augmentation approach with the standard one, i.e. augmenting the time-series in the input space. We show that augmentation methods on the feature space leads to 20% increase in risk-adjusted return compared to a model trained with transfer learning but without augmentation.

Foramitti, J. (2021). “[AgentPy: A package for agent-based modeling in Python.](#)” In: *Journal of Open Source Software* 6(62), p. 3065.

AgentPy is an open-source library for the development and analysis of agent-based models.

Frahm, G. (2015). “[A theoretical foundation of portfolio resampling.](#)” In: *Theory and Decision* 79(1), pp. 107–132.

A portfolio-resampling procedure invented by Richard and Robert Michaud is a subject of highly controversial discussion and big scientific dispute. It has been evaluated in many empirical studies and Monte Carlo experiments. Apart from the contradictory findings, the Michaud approach still lacks a theoretical foundation. I prove

that portfolio resampling has a strong foundation in the classic theory of rational behavior. Every noise trader could do better by applying the Michaud procedure. By contrast, a signal trader who has enough prediction power and risk-management skills should refrain from portfolio resampling. The key note is that in most simulation studies, investors are considered as noise traders. This explains why portfolio resampling performs well in simulation studies, but could be mediocre in real life.

- Franco-Pedroso, J., Gonzalez-Rodriguez, J., Planas, M., Cubero, J., Cobo, R., and Pablos, F. (2019a). “Generating Virtual Scenarios of Multivariate Financial Data for Quantitative Trading Applications.” In: *The Journal of Financial Data Science* 1 (2), pp. 55–77.

In this article, the authors present a novel approach to the generation of virtual scenarios of multivariate financial data of arbitrary length and composition of assets. With this approach, decades of realistic time synchronized data can be simulated for a large number of assets, producing diverse scenarios to test and improve quantitative investment strategies. The authors’ approach is based on the analysis and synthesis of the time dependent individual and joint characteristics of real financial time series, using stochastic sequences of market trends to draw multivariate returns from time-dependent probability functions that preserve both distributional properties of asset returns and time-dependent correlation among time series. Moreover, new time-synchronized assets can be arbitrarily generated through a principal component analysis-based procedure to obtain any number of assets in the final virtual scenario. The validation of such a simulation is tested with an extensive set of measurements and shows a significant degree of agreement with the reference performance of real financial series, better than that obtained with other classical and state-of-the-art approaches.

- Franco-Pedroso, J., Gonzalez-Rodriguez, J., Planas, M., Cubero, J., Cobo, R., and Pablos, F. (2019b). “The ETS challenges: a machine learning approach to the evaluation of simulated financial time series for improving generation processes.” In: *The Journal of Financial Data Science* 1(3) (3), pp. 68–86.

This paper presents an evaluation framework that attempts to quantify the “degree of realism” of simulated financial time series, whatever the simulation method could be, with the aim of discover unknown characteristics that are not being properly reproduced by such methods in order to improve them. For that purpose, the evaluation framework is posed as a machine learning problem in which some given time series examples have to be classified as simulated or real financial time series. The “challenge” is proposed as an open competition, similar to those published at the Kaggle platform, in which participants must send their classification results along with a description of the features and the classifiers used. The results of these “challenges” have revealed some interesting properties of financial data, and have lead to substantial improvements in our simulation methods under research, some of which will be described in this work.

- Franses, P. H. and Bruijn, B. de (2017). “Benchmarking Judgmentally Adjusted Forecasts.” In: *International Journal of Finance and Economics* 22(1), pp. 3–11.

Many publicly available macroeconomic forecasts are judgmentally adjusted model-based forecasts. In practice, usually only a single final forecast is available, and not the underlying econometric model, nor are the size and reason for adjustment known. Hence, the relative weights given to the model forecasts and to the judgement are usually unknown to the analyst. This paper proposes a methodology to evaluate the quality of such final forecasts, also to allow learning from past errors. To do so, the analyst needs benchmark forecasts. We propose two such benchmarks. The first is the simple no-change forecast, which is the bottom line forecast that an expert should be able to improve. The second benchmark is an estimated model-based forecast, which is found as the best forecast given the realizations and the final forecasts. We illustrate this methodology for two sets of GDP growth forecasts, one for the USA and one for the Netherlands. These applications tell us that adjustment appears most effective in periods of first recovery from a recession.

- Fritsch, S., Timphus, M., and Weiss, G. N. F. (2021). “Marginals Versus Copulas: Which Account For More Model Risk In Multivariate Risk Forecasting?” In: *SSRN e-Print*.

Copulas. We study the model risk of multivariate risk models in a comprehensive empirical study on Copula-GARCH models used for forecasting Value-at-Risk and Expected Shortfall. To determine whether model risk inherent in the forecasting of portfolio risk is caused by the candidate marginal or copula models, we analyze different groups of models in which we fix either the marginals, the copula, or neither. Model risk is economically significant, is especially high during periods of crisis, and is almost completely due to the choice of the copula. We then propose the use of the model confidence set procedure to narrow down the set of available models and reduce model risk for Copula-GARCH risk models. Our proposed approach leads to a significant improvement in the mean absolute deviation of one day ahead forecasts by our various candidate risk models.

Fu, B., Kirchbuchner, F., and Kuijper, A. (2020). “Data augmentation for time series.” In: *Proceedings of the 13th ACM International Conference on Pervasive Technologies Related to Assistive Environments*. ACM.

Large labeled quantities and diversities of training data are often needed for supervised, data-based modelling. Data distribution should cover a rich representation to support the generalizability of the trained end-to-end inference model. However, this is often hindered by limited labeled data and the expensive data collection process, especially for human activity recognition tasks. Extensive manual labeling is required. Data augmentation is thus a widely used regularization method for deep learning, especially applied on image data to increase the classification accuracy. But it is less researched for time series. In this paper, we investigate the data augmentation task on continuous capacitive time series with the example on exercise recognition. We show that the traditional data augmentation can enrich the source distribution and thus make the trained inference model more generalized. This further increases the recognition performance for unseen target data around 21.4 percentage points compared to inference model without data augmentation. The generative models such as variational autoencoder or conditional variational autoencoder can further reduce the variance on the target data.

Fulle, M. J. and Herwartz, H. (2021). “A Multivariate Markov-Switching GARCH Model with Copula-Distributed Innovations.” In: *SSRN e-Print*.

In order to improve the dynamic assessment of financial market interdependencies, we develop a new Markov switching approach to multivariate volatility modelling. More specific, we take advantage of the flexible copula multivariate GARCH model of Lee and Long (2009), and allow state dependence with regard to the dependence structure governed by copula functions. We show some asymptotic features of the new model and motivate a two-step approach to maximum likelihood (ML)-estimation. As an empirical illustration, we consider a bivariate series comprising the returns of a high-yield equity index (MSCI World Developed Markets) and a traditional safe-haven asset (gold), and compare the goodness-of-fit of the new model with existing multivariate GARCH approaches. The new model offers advantages when it comes to detect and depict flows between financial assets in order to reduce portfolio risks in extreme market situations. We find that the suggested Markov switching generalization of the copula-MGARCH model of Lee and Long (2009) benefits from capturing structural changes in the non-linear dependence patterns, and helps to uncover flight-to-safety effects in times of economic turmoil.

Galeeva, R., Hoogland, J., and Eydeland, A. (2012). “Measuring correlation risk for Energy Derivatives.” In: *Risk Management in Commodity Markets: From Shipping to Agriculturals and Energy*.

This chapter discusses 1) Correlation 2) Perturbing the Correlation Matrix 3) Correlation VaR.

Gao, G., Mishra, B., and Ramazzotti, D. (2017). “Efficient Simulation of Financial Stress Testing Scenarios with Suppes-Bayes Causal Networks.” In: *arXiv e-Print*.

The most recent financial upheavals have cast doubt on the adequacy of some of the conventional quantitative risk management strategies, such as VaR (Value at Risk), in many common situations. Consequently, there has been an increasing need for verisimilar financial stress testings, namely simulating and analyzing financial portfolios in extreme, albeit rare scenarios. Unlike conventional risk management which exploits statistical correlations among financial instruments, here we focus our analysis on the notion of probabilistic causation, which is embodied by Suppes-Bayes Causal Networks (SBCNs), SBCNs are probabilistic graphical models that have many attractive features in terms of more accurate causal analysis for generating financial stress scenarios. In this paper, we present a novel approach for conducting stress testing of financial portfolios based on SBCNs in combination with classical machine learning classification tools. The resulting method is shown to be capable of correctly discovering the causal relationships among financial factors that affect the portfolios and thus, simulating stress testing scenarios with a higher accuracy and lower computational complexity than conventional Monte Carlo Simulations.

Gao, G., Mishra, B., and Ramazzotti, D. (2018). “Causal data science for financial stress testing.” In: *Journal of Computational Science* 26, pp. 294–304.

The most recent financial upheavals have cast doubt on the adequacy of some of the conventional quantitative risk management strategies, such as VaR (Value at Risk), in many common situations. Consequently, there has been an increasing need for verisimilar financial stress testings, namely simulating and analyzing financial portfolios in extreme, albeit rare scenarios. Unlike conventional risk management which exploits statistical correlations among financial instruments, here we focus our analysis on the notion of probabilistic causation, which is embodied by Suppes-Bayes Causal Networks (SBCNs); SBCNs are probabilistic graphical models that have many attractive features in terms of more accurate causal analysis for generating financial stress scenarios. In this paper, we present a novel approach for conducting stress testing of financial portfolios based on SBCNs in combination with classical machine learning classification tools. The resulting method is shown to be capable of correctly

discovering the causal relationships among financial factors that affect the portfolios and thus, simulating stress testing scenarios with a higher accuracy and lower computational complexity than conventional Monte Carlo Simulations.

- Gharghabi, S., Imani, S., Bagnall, A., Darvishzadeh, A., and Keogh, E. (2018). “Matrix profile XII: mpdist: A novel time series distance measure to allow data mining in more challenging scenarios.” In: *IEEE International Conference on Data Mining (ICDM)*. IEEE, pp. 965–970.

At their core, many time series data mining algorithms can be reduced to reasoning about the shapes of time series subsequences. This requires a distance measure, and most algorithms use Euclidean Distance or Dynamic Time Warping (DTW) as their core subroutine. We argue that these distance measures are not as robust as the community believes. The undue faith in these measures derives from an overreliance on benchmark datasets and self-selection bias. The community is reluctant to address more difficult domains, for which current distance measures are ill-suited. In this work, we introduce a novel distance measure MPdist. We show that our proposed distance measure is much more robust than current distance measures. Furthermore, it allows us to successfully mine datasets that would defeat any Euclidean or DTW distance-based algorithm. Additionally, we show that our distance measure can be computed so efficiently, it allows analytics on fast streams.

- Gilleland, E. (2020a). “Bootstrap Methods for Statistical Inference. Part I: Comparative Forecast Verification for Continuous Variables.” In: *Journal of Atmospheric and Oceanic Technology* 37(11), pp. 2117–2134.

When making statistical inferences, bootstrap resampling methods are often appealing because of less stringent assumptions about the distribution of the statistic(s) of interest. However, the procedures are not free of assumptions. This paper addresses a specific situation that occurs frequently in atmospheric sciences where the standard bootstrap is not appropriate: comparative forecast verification of continuous variables. In this setting, the question to be answered concerns which of two weather or climate models is better in the sense of some type of average deviation from observations. The series to be compared are generally strongly dependent, which invalidates the most basic bootstrap technique. This paper also introduces new bootstrap code from the R package “distillery” that facilitates easy implementation of appropriate methods for paired-difference-of-means bootstrap procedures for dependent data.

- Gilleland, E. (2020b). “Bootstrap Methods for Statistical Inference. Part II: Extreme-Value Analysis.” In: *Journal of Atmospheric and Oceanic Technology* 37(11), pp. 2135–2144.

This paper is the sequel to a companion paper on bootstrap resampling that reviews bootstrap methodology for making statistical inferences for atmospheric science applications where the necessary assumptions are often not met for the most commonly used resampling procedures. In particular, this sequel addresses extreme-value analysis applications with discussion on the challenges for finding accurate bootstrap methods in this context. New bootstrap code from the R packages “distillery” and “extRemes” is introduced. It is further found that one approach for accurate confidence intervals in this setting is not well suited to the case when the random sample’s distribution is not stationary.

- Glasserman, P., Kang, C., and Kang, W. (2015). “Stress scenario selection by empirical likelihood.” In: *Quantitative Finance* 15(1), pp. 25–41.

This paper develops a method for selecting and analysing stress scenarios for financial risk assessment, with particular emphasis on identifying sensible combinations of stresses to multiple factors. We focus primarily on reverse stress testing - finding the most likely scenarios leading to losses exceeding a given threshold. We approach this problem using a nonparametric empirical likelihood estimator of the conditional mean of the underlying market factors given large losses. We then scale confidence regions for the conditional mean by a coefficient that depends on the tails of the market factors to estimate the most likely loss scenarios. We provide rigorous justification for the confidence regions and the scaling procedure when the joint distribution of the market factors and portfolio loss is elliptically contoured. We explicitly characterize the impact of the heaviness of the tails of the distribution, contrasting a broad spectrum of cases including exponential tails and regularly varying tails. The key to this analysis lies in the asymptotics of the conditional variances and covariances in extremes. These results also lead to asymptotics for marginal expected shortfall and the corresponding variance, conditional on a market stress; we combine these results with empirical likelihood significance tests of systemic risk rankings based on marginal expected shortfall in stress scenarios.

- Goel, K., Gu, A., Li, Y., and Re, C. (2020). “Model Patching: Closing the Subgroup Performance Gap with Data Augmentation.” In: *arXiv e-Print*.

Classifiers in machine learning are often brittle when deployed. Particularly concerning are models with inconsistent performance on specific subgroups of a class, e.g., exhibiting disparities in skin cancer classification in



the presence or absence of a spurious bandage. To mitigate these performance differences, we introduce model patching, a two-stage framework for improving robustness that encourages the model to be invariant to subgroup differences, and focus on class information shared by subgroups. Model patching first models subgroup features within a class and learns semantic transformations between them, and then trains a classifier with data augmentations that deliberately manipulate subgroup features. We instantiate model patching with CAMEL, which (1) uses a CycleGAN to learn the intra-class, inter-subgroup augmentations, and (2) balances subgroup performance using a theoretically-motivated subgroup consistency regularizer, accompanied by a new robust objective. We demonstrate CAMEL’s effectiveness on 3 benchmark datasets, with reductions in robust error of up to 33% relative to the best baseline. Lastly, CAMEL successfully patches a model that fails due to spurious features on a real-world skin cancer dataset.

Golub, B., Greenberg, D., and Ratcliffe, R. (2018). “Market-Driven Scenarios: An Approach for Plausible Scenario Construction.” In: *The Journal of Portfolio Management* 44(5), pp. 6–20.

The use of scenario analysis to better understand portfolios has increased significantly since the global financial crisis. In this article, the authors describe a stress scenario framework and process that has been developed for risk management and investment management purposes. This hybrid framework, which the authors refer to as market-driven scenarios, works as follows. Scenario forecasts of key market indicators are first formulated by market practitioners. An econometric framework then uses these indicators as inputs to imply plausible shocks to a global set of risk factors. These factor shocks are finally put into a portfolio valuation engine, yielding hypothetical fund profit and loss (P&L) that can be decomposed into its underlying drivers. Key to the effectiveness of this approach is the cross-functional involvement of investors, risk managers, and economists. In conjunction, the authors define potential geopolitical or other macro events, specify potential economic outcomes, and translate them into shocks to key policy risk variables and risk model factors. The process is completed by applying the shocks to portfolios and evaluating whether P&L outcomes are consistent with fund mandates and whether positioning is deliberate, diversified, and scaled.

Goncalves, S. and Perron, B. (2020). “Bootstrapping factor models with cross sectional dependence.” In: *Journal of Econometrics* 218(2), pp. 476–495.

We consider bootstrap methods for factor-augmented regressions with cross sectional dependence among idiosyncratic errors. This is important to capture the bias of the OLS estimator derived recently by Goncalves and Perron (2014). We first show that a common approach of resampling cross sectional vectors over time is invalid in this context because it induces a zero bias. We then propose the cross-sectional dependent (CSD) bootstrap where bootstrap samples are obtained by taking a random vector and multiplying it by the square root of a consistent estimator of the covariance matrix of the idiosyncratic errors. We show that if the covariance matrix estimator is consistent in the spectral norm, then the CSD bootstrap is consistent, and we verify this condition for the thresholding estimator of Bickel and Levina (2008). Finally, we apply our new bootstrap procedure to forecasting inflation using convenience yields as recently explored by Gospodinov and Ng (2013).

Gonzalez, D. P., Rebollo, F. F., Polo, F. J. G., and Vyetrenko, S. (2021). “Similarity Metrics for Transfer Learning in Financial Markets.” In: *ICAPS Workshop on Planning for Financial Services (FinPlan 2021)*.

Markov Decision Processes (MDPs) are an effective way to formally describe many Machine Learning problems. In fact, recently MDPs have also emerged as a powerful framework to model financial trading tasks. For example, financial MDPs can model different market scenarios. However, the learning of a (near-)optimal policy for each of these financial MDPs can be a very time-consuming process, especially when nothing is known about the policy to begin with. An alternative approach is to find a similar financial MDP for which we have already learned its policy, and then reuse such policy in the learning of a new policy for a new financial MDP. Such a knowledge transfer between market scenarios raises several issues. On the one hand, how to measure the similarity between financial MDPs. On the other hand, how to use this similarity measurement to effectively transfer the knowledge between financial MDPs. This paper addresses both of these issues. Regarding the first one, this paper analyzes the use of three similarity metrics based on conceptual, structural and performance aspects of the financial MDPs. Regarding the second one, this paper uses Probabilistic Policy Reuse to balance the exploitation/exploration in the learning of a new financial MDP according to the similarity of the previous financial MDPs whose knowledge is reused.

Großer, J. and Okhrin, O. (2021). “Copulae: An overview and recent developments.” In: *WIREs Computational Statistics*.

Over the decades that have passed since they were introduced, copulae still remain a very powerful tool for modeling and estimating multivariate distributions. This work gives an overview of copula theory and it also



summarizes the latest results. This article recalls the basic definition, the most important cases of bivariate copulae, and it then proceeds to a sketch of how multivariate copulae are developed both from bivariate copulae and from scratch. Regarding higher dimensions, the focus is on hierarchical Archimedean, vine, and factor copulae, which are the most often used and most flexible ways to introduce copulae to multivariate distributions. We also provide an overview of how copulae can be used in various fields of data science, including recent results. These fields include but are not limited to time series and machine learning. Finally, we describe estimation and testing methods for copulae in general, their application to the presented copula structures, and we give some specific testing and estimation procedures for those specific copulae.

Grundke, P. (2012). “Further recipes for quantitative reverse stress testing.” In: *Journal of Risk Model Validation* 6(2).

As a consequence of the 2007-9 financial crisis, the regulatory authorities now oblige banks to carry out reverse stress tests. This new instrument in the stress test toolbox aims to find those scenarios that cause a bank to cross the frontier between survival and default. Afterward, the scenario that is most likely to occur has to be identified. In a 2011 paper by Grundke it is argued that bottom-up approaches, which are a specific integrated risk measurement technique, are basically well suited to solving the inversion problem inherent in a reverse stress test and, in particular, for computing the probabilities of reverse stress test scenarios while taking existing risk dependencies into account. Building upon the earlier work of Grundke, this paper shows how the modeling framework previously presented can be extended to incorporate more real-world features such as a time-varying credit quality of the bank or contagion effects. The consequences on the results of the reverse stress test are analyzed for each modification.

Gzyl, H., Horst, E. ter, and Molina, G. (2017). “Inferring probability densities from expert opinion.” In: *Applied Mathematical Modelling* 43, pp. 306–320.

When experts are asked to assess a situation, they often express their opinions providing estimates of the probability of observing the occurrence of a random variable in given intervals, sometimes up to a range of values, rather than simply providing point estimates. The problem we face is how to translate that expert opinion into probability distributions. We examine a novel way of solving that problem by making use of the maximum entropy method in the data to deal with expert opinions expressed with or without uncertainty bands. Our method allows us to unveil underlying probability distributions driving expert opinions expressed with and without uncertainty.

Haldane, A. G. and Turrell, A. E. (2018). “Drawing on different disciplines: macroeconomic agent-based models.” In: *Journal of Evolutionary Economics* 29(1), pp. 1–28.

Macroeconomic modelling has been under intense scrutiny since the Great Financial Crisis, when serious shortcomings were exposed in the methodology used to understand the economy as a whole. Criticism has been levelled at the assumptions employed in the dominant models, particularly that economic agents are homogeneous and optimising and that the economy is equilibrating. In a related paper (Haldane and Turrell Oxford Rev Econ Polic 34(1-2):219-251 2018), we argue that an interdisciplinary approach to modelling in macroeconomics is beneficial. Here we focus on what one such approach - agent-based modelling, which has been extensively used across a wide range of disciplines - could do for macroeconomics. Agent-based models are complementary to existing approaches to macroeconomics and are particularly well-suited to answering questions where complexity, heterogeneity, networks, and heuristics play an important role.

Hambuckers, J. and Heuchenne, C. (2016). “Estimating the Out-of-Sample Predictive Ability of Trading Rules: A Robust Bootstrap Approach.” In: *Journal of Forecasting* 35(4), pp. 347–372.

In this paper, we provide a novel way to estimate the out-of-sample predictive ability of a trading rule. Usually, this ability is estimated using a sample-splitting scheme, true out-of-sample data being rarely available. We argue that this method makes poor use of the available data and creates data-mining possibilities. Instead, we introduce an alternative bootstrap approach. This method enables building in-sample and out-of-sample bootstrap datasets that do not overlap but exhibit the same time dependencies. We show in a simulation study that this technique drastically reduces the mean squared error of the estimated predictive ability. We illustrate our methodology on IBM, MSFT and DJIA stock prices, where we compare 11 trading rules specifications. For the considered datasets, two different filter rule specifications have the highest out-of-sample mean excess returns. However, all tested rules cannot beat a simple buy-and-hold strategy when trading at a daily frequency.

Harutyunyan, H., Moyer, D., Khachatryan, H., Steeg, G. V., and Galstyan, A. (2021). “Efficient Covariance Estimation from Temporal Data.” In: *arXiv e-Print*.

Estimating the covariance structure of multivariate time series is a fundamental problem with a wide-range of real-world applications – from financial modeling to fMRI analysis. Despite significant recent advances, current state-of-the-art methods are still severely limited in terms of scalability, and do not work well in high-dimensional undersampled regimes. In this work we propose a novel method called Temporal Correlation Explanation, or T-CorEx, that (a) has linear time and memory complexity with respect to the number of variables, and can scale to very large temporal datasets that are not tractable with existing methods; (b) gives state-of-the-art results in highly undersampled regimes on both synthetic and real-world datasets; and (c) makes minimal assumptions about the character of the dynamics of the system. T-CorEx optimizes an information-theoretic objective function to learn a latent factor graphical model for each time period and applies two regularization techniques to induce temporal consistency of estimates. We perform extensive evaluation of T-CorEx using both synthetic and real-world data and demonstrate that it can be used for detecting sudden changes in the underlying covariance matrix, capturing transient correlations and analyzing extremely high-dimensional complex multivariate time series such as high-resolution fMRI data.

Harwick, C. (2021). “[Helipad: A Framework for Agent-Based Modeling in Python.](#)” In: *SSRN e-Print*.

Agent-based modeling tools commonly trade off usability against power and vice versa. On the one hand, full development environments like NetLogo feature a shallow learning curve, but have a relatively limited proprietary language. Others written in Python or Matlab, for example, have the advantage of a full-featured language with a robust community of third-party libraries, but are typically more skeletal and require more setup and boilerplate in order to write a model. Helipad is introduced to fill this gap. Helipad is a new agent-based modeling framework for Python with the goal of a shallow learning curve, extensive flexibility, minimal boilerplate, and powerful yet easy to set up visualization, in a full Python environment. We summarize Helipad’s general architecture and capabilities, and briefly preview a variety of models from a variety of disciplines, including multilevel models, matching models, network models, spatial models, and others.

He, Z., Xie, L., Chen, X., Zhang, Y., Wang, Y., and Tian, Q. (2019). “[Data Augmentation Revisited: Rethinking the Distribution Gap between Clean and Augmented Data.](#)” In: *arXiv e-Print*.

Data augmentation has been widely applied as an effective methodology to improve generalization in particular when training deep neural networks. Recently, researchers proposed a few intensive data augmentation techniques, which indeed improved accuracy, yet we notice that these methods augment data have also caused a considerable gap between clean and augmented data. In this paper, we revisit this problem from an analytical perspective, for which we estimate the upper-bound of expected risk using two terms, namely, empirical risk and generalization error, respectively. We develop an understanding of data augmentation as regularization, which highlights the major features. As a result, data augmentation significantly reduces the generalization error, but meanwhile leads to a slightly higher empirical risk. On the assumption that data augmentation helps models converge to a better region, the model can benefit from a lower empirical risk achieved by a simple method, i.e., using less-augmented data to refine the model trained on fully-augmented data. Our approach achieves consistent accuracy gain on a few standard image classification benchmarks, and the gain transfers to object detection.

Heaton, J. B. and Witte, J. (2021). “[Synthetic Financial Data: An Application to Regulatory Compliance for Broker-Dealers.](#)” In: *SSRN e-Print*.

Big Data hype has not missed investment management, the reality is that price data from U.S. financial markets are not really Big Data. Price data is Small Data. The fact that sellers and advisors in financial markets use Small Data to generate and test investment strategies creates two major problems. First, the economic mechanisms that generate prices (and therefore returns) may change through time, so that historical data from an earlier time may tell us little or nothing about future prices and returns. Second, even if data-generating-mechanisms are somewhat stable through time, inferences about the profitability of investment strategies may be sensitive to a handful of outliers in the data that get picked up again and again in different strategies mined from the same Small Data set. In this article, we present an answer to the financial Small Data problem: using machine-learning methods to generate “synthetic” financial data. The essential part of our approach to developing synthetic data is the use of machine learning methods to generate data that might have been generated by financial markets but was not. Synthetic price and return data have numerous uses, including testing new investment strategies and helping investors plan for retirement and other personal investment goals with more realistic future return scenarios. In this article, we focus on a particularly important use of synthetic data: meeting legal and regulatory requirements such as best interest and fiduciary requirements.

Henry-Labordere, P. (2019). “[Generative models for financial data.](#)” In: *SSRN e-Print*.

- In this paper, by relying on a new algorithm for computing efficiently Wasserstein distance, we construct non-parametric generative models. As some applications, we consider automatic generation of financial data and anomaly detection. The efficiency of our algorithm is compared against GAN and Wasserstein-GAN methods.
- Hofert, M., Prasad, A., and Zhu, M. (2021). “Multivariate time-series modeling with generative neural networks.” In: *arXiv e-Print*.
- Generative moment matching networks (GMMNs) are introduced as dependence models for the joint innovation distribution of multivariate time series (MTS). Following the popular copula-GARCH approach for modeling dependent MTS data, a framework based on a GMMN-GARCH approach is presented. First, ARMA-GARCH models are utilized to capture the serial dependence within each univariate marginal time series. Second, if the number of marginal time series is large, principal component analysis (PCA) is used as a dimension-reduction step. Last, the remaining cross-sectional dependence is modeled via a GMMN, the main contribution of this work. GMMNs are highly flexible and easy to simulate from, which is a major advantage over the copula-GARCH approach. Applications involving yield curve modeling and the analysis of foreign exchange rate returns demonstrate the utility of the GMMN-GARCH approach, especially in terms of producing better empirical predictive distributions and making better probabilistic forecasts. All results are reproducible with the demo GMMN\_MTS\_paper of the R package gnn.
- Hoffman, J. I. E. (2015). “Resampling Statistics.” In: *Biostatistics for Medical and Biomedical Practitioners*. Elsevier, pp. 655–661.
- It is possible to ignore the type of distribution and still obtain valid estimates of means, standard deviations, slopes, or confidence limits that sometimes cannot be obtained in any other way, especially if the sample size is small and not repeatable. This is achieved by repeated sampling of randomly chosen data from the single data set. The methods are computer intensive but this is no longer a barrier to implementing them.
- Hoffmann, J., Bar-Sinai, Y., Lee, L., Andrejevic, J., Mishra, S., Rubinstein, S. M., and Rycroft, C. H. (2018). “Machine Learning in a data-limited regime: Augmenting experiments with synthetic data uncovers order in crumpled sheets.” In: *arXiv e-Print*.
- Machine learning has gained widespread attention as a powerful tool to identify structure in complex, high-dimensional data. However, these techniques are ostensibly inapplicable for experimental systems with limited data acquisition rates, due to the restricted size of the dataset. Here we introduce a strategy to resolve this impasse by augmenting the experimental dataset with synthetically generated data of a much simpler sister system. Specifically, we study spontaneously emerging local order in crease networks of crumpled thin sheets, a paradigmatic example of spatial complexity, and show that machine learning techniques can be effective even in a data-limited regime. This is achieved by augmenting the scarce experimental dataset with inexhaustible amounts of simulated data of flat-folded sheets, which are simple to simulate. This significantly improves the predictive power in a test problem of pattern completion and demonstrates the usefulness of machine learning in bench-top experiments where data is good but scarce.
- Honore, B. E. and Hu, L. (2017). “Poor (Wo)man’s Bootstrap.” In: *Econometrica* 85(4), pp. 1277–1301.
- The bootstrap is a convenient tool for calculating standard errors of the parameter estimates of complicated econometric models. Unfortunately, the fact that these models are complicated often makes the bootstrap extremely slow or even practically infeasible. This paper proposes an alternative to the bootstrap that relies only on the estimation of one-dimensional parameters. We introduce the idea in the context of M and GMM estimators. A modification of the approach can be used to estimate the variance of two-step estimators.
- Horowitz, J. L. (2019). “Bootstrap methods in econometrics.” In: *Annual review of economics* 11(1), pp. 193–224.
- The bootstrap is a method for estimating the distribution of an estimator or test statistic by resampling one’s data or a model estimated from the data. Under conditions that hold in a wide variety of econometric applications, the bootstrap provides approximations to distributions of statistics, coverage probabilities of confidence intervals, and rejection probabilities of hypothesis tests that are more accurate than the approximations of first-order asymptotic distribution theory. The reductions in the differences between true and nominal coverage or rejection probabilities can be very large. In addition, the bootstrap provides a way to carry out inference in certain settings where obtaining analytic distributional approximations is difficult or impossible. This article explains the usefulness and limitations of the bootstrap in contexts of interest in econometrics. The presentation is informal and expository. It provides an intuitive understanding of how the bootstrap works. Mathematical details are available in the references that are cited.
- Hsu, M.-W., Lessmann, S., Sung, M.-C., Ma, T., and Johnson, J. E. V. (2016). “Bridging the divide in financial

market forecasting: machine learners vs. financial economists.” In: *Expert Systems with Applications* 61, pp. 215–234.

An extensive benchmark in financial time series forecasting is performed. Best machine learning (ML) methods out-perform best econometric methods. The ML methodology employed significantly affects forecasting accuracy. Market maturity, forecast horizon and model-assessment method affect forecast accuracy. Evidence against the informational value of technical indicators. Financial time series forecasting is a popular application of machine learning methods. Previous studies report that advanced forecasting methods predict price changes in financial markets with high accuracy and that profit can be made trading on these predictions. However, financial economists point to the informational efficiency of financial markets, which questions price predictability and opportunities for profitable trading. The objective of the paper is to resolve this contradiction. To this end, we undertake an extensive forecasting simulation, based on data from thirty-four financial indices over six years. These simulations confirm that the best machine learning methods produce more accurate forecasts than the best econometric methods. We also examine the methodological factors that impact the predictive accuracy of machine learning forecasting experiments. The results suggest that the predictability of a financial market and the feasibility of profitable model-based trading are significantly influenced by the maturity of the market, the forecasting method employed, the horizon for which it generates predictions and the methodology used to assess the model and simulate model-based trading. We also find evidence against the informational value of indicators from the field of technical analysis. Overall, we confirm that advanced forecasting methods can be used to predict price changes in some financial markets and we discuss whether these results question the prevailing view in the financial economics literature that financial markets are efficient.

Hu, J., Akande, O., and Wang, Q. (2021). “Multiple Imputation and Synthetic Data Generation with the R package NPBayesImputeCat.” In: *arXiv e-Print*.

In many contexts, missing data and disclosure control are ubiquitous and challenging issues. In particular at statistical agencies, the respondent-level data they collect from surveys and censuses can suffer from high rates of missingness. Furthermore, agencies are obliged to protect respondents’ privacy when publishing the collected data for public use. The NPBayesImputeCat R package, introduced in this paper, provides routines to i) create multiple imputations for missing data, and ii) create synthetic data for statistical disclosure control, for multivariate categorical data, with or without structural zeros. We describe the Dirichlet process mixture of products of multinomial distributions model used in the package, and illustrate various uses of the package using data samples from the American Community Survey (ACS). We also compare results of the missing data imputation to the mice R package and those of the synthetic data generation to the synthpop R package.

Huang, M. and Yu, S. (2020). “A new procedure for resampled portfolio with shrinkaged covariance matrix.” In: *Journal of Applied Statistics* 47(44), pp. 642–652.

Dealing with estimation error is an important issue when we implement the mean-variance paradigm for portfolio construction. To tackle the problem, two approaches are proposed in literature, the portfolio resampling technique introduced by Michaud and the well-known shrinkaged covariance matrix method. There are certain evidences on the advantages of shrinkaged covariance over portfolio resampling, however, it is unclear whether a combination of the two approaches could produce a better performance compared with using shrinkaged covariance alone. In this paper, we propose a new algorithm to integrated linear or nonlinear shrinkage estimation with resampled portfolio to achieve a further improvement. Our method are demonstrated via extensive simulation and application in active portfolio management process.

Imbens, G. and Menzel, K. (2021). “A causal bootstrap.” In: *Annals of Statistics* 49(3).

The bootstrap, introduced by The Jackknife, the Bootstrap and Other Resampling Plans ((1982), SIAM), has become a very popular method for estimating variances and constructing confidence intervals. A key insight is that one can approximate the properties of estimators by using the empirical distribution function of the sample as an approximation for the true distribution function. This approach views the uncertainty in the estimator as coming exclusively from sampling uncertainty. We argue that for causal estimands the uncertainty arises entirely, or partially, from a different source, corresponding to the stochastic nature of the treatment received. We develop a bootstrap procedure for inference regarding the average treatment effect that accounts for this uncertainty, and compare its properties to that of the classical bootstrap. We consider completely randomized and observational designs as well as designs with imperfect compliance.

Imokoyende, M. (2019). “Variational Autoencoder In Finance.” In: *Towards Data Science*.

Istrate, G. (2021). “Models we Can Trust: Toward a Systematic Discipline of (Agent-Based) Model Interpretation and Validation.” In: *arXiv e-Print*.

We advocate the development of a discipline of interacting with and extracting information from models, both mathematical (e.g. game-theoretic ones) and computational (e.g. agent-based models). We outline some directions for the development of a such a discipline: - the development of logical frameworks for the systematic formal specification of stylized facts and social mechanisms in (mathematical and computational) social science. Such frameworks would bring to attention new issues, such as phase transitions, i.e. dramatical changes in the validity of the stylized facts beyond some critical values in parameter space. We argue that such statements are useful for those logical frameworks describing properties of ABM. - the adaptation of tools from the theory of reactive systems (such as bisimulation) to obtain practically relevant notions of two systems "having the same behavior". - the systematic development of an adversarial theory of model perturbations, that investigates the robustness of conclusions derived from models of social behavior to variations in several features of the social dynamics. These may include: activation order, the underlying social network, individual agent behavior.

Iwana, B. K. and Uchida, S. (2021). "An empirical survey of data augmentation for time series classification with neural networks." In: *PLOS One* 16(7). Ed. by F. Schwenker, e0254841.

In recent times, deep artificial neural networks have achieved many successes in pattern recognition. Part of this success can be attributed to the reliance on big data to increase generalization. However, in the field of time series recognition, many datasets are often very small. One method of addressing this problem is through the use of data augmentation. In this paper, we survey data augmentation techniques for time series and their application to time series classification with neural networks. We propose a taxonomy and outline the four families in time series data augmentation, including transformation-based methods, pattern mixing, generative models, and decomposition methods. Furthermore, we empirically evaluate 12 time series data augmentation methods on 128 time series classification datasets with six different types of neural networks. Through the results, we are able to analyze the characteristics, advantages and disadvantages, and recommendations of each data augmentation method. This survey aims to help in the selection of time series data augmentation for neural network applications.

Jabbar, A., Li, X., and Omar, B. (2020). "A Survey on Generative Adversarial Networks: Variants, Applications, and Training." In: *arXiv e-Print*.

The Generative Models have gained considerable attention in the field of unsupervised learning via a new and practical framework called Generative Adversarial Networks (GAN) due to its outstanding data generation capability. Many models of GAN have proposed, and several practical applications emerged in various domains of computer vision and machine learning. Despite GAN's excellent success, there are still obstacles to stable training. The problems are due to Nash-equilibrium, internal covariate shift, mode collapse, vanishing gradient, and lack of proper evaluation metrics. Therefore, stable training is a crucial issue in different applications for the success of GAN. Herein, we survey several training solutions proposed by different researchers to stabilize GAN training. We survey, (I) the original GAN model and its modified classical versions, (II) detail analysis of various GAN applications in different domains, (III) detail study about the various GAN training obstacles as well as training solutions. Finally, we discuss several new issues as well as research outlines to the topic.

Jaeger, M., Krugel, S., Marinelli, D., Papenbrock, J., and Schwendner, P. (2020). "Understanding machine learning for diversified portfolio construction by explainable AI." In: *SSRN e-Print*.

In this paper, we construct a pipeline to investigate heuristic diversification strategies in asset allocation. We use machine learning concepts ("explainable AI") to compare the robustness of different strategies and back out implicit rules for decision making. In a first step, we augment the asset universe (the empirical dataset) with a range of scenarios generated with a block bootstrap from the empirical dataset. Second, we backtest the candidate strategies over a long period of time, checking their performance variability. Third, we use XGBoost as a regression model to connect the difference between the measured performances between two strategies to a pool of statistical features of the portfolio universe tailored to the investigated strategy. Finally, we employ the concept of Shapley values to extract the relationships that the model could identify between the portfolio characteristics and the statistical properties of the asset universe. We test this pipeline for studying risk-parity strategies with a volatility target, and in particular, comparing the machine learning-driven Hierarchical Risk Parity (HRP) to the classical Equal Risk Contribution (ERC) strategy. In the augmented dataset built from a multi-asset investment universe of commodities, equities and fixed income futures, we find that HRP better matches the volatility target, and shows better risk-adjusted performances. Finally, we train XGBoost to learn the difference between the realized Calmar ratios of HRP and ERC and extract explanations. The explanations provide fruitful ex-post indications of the connection between the statistical properties of the universe and



the strategy performance in the training set. For example, the model confirms that features addressing the hierarchical properties of the universe are connected to the relative performance of HRP respect to ERC.

Janke, T., Ghanmi, M., and Steinke, F. (2021). “Implicit Generative Copulas.” In: *arXiv e-Print*.

Copulas are a powerful tool for modeling multivariate distributions as they allow to separately estimate the univariate marginal distributions and the joint dependency structure. However, known parametric copulas offer limited flexibility especially in high dimensions, while commonly used non-parametric methods suffer from the curse of dimensionality. A popular remedy is to construct a tree-based hierarchy of conditional bivariate copulas. In this paper, we propose a flexible, yet conceptually simple alternative based on implicit generative neural networks. The key challenge is to ensure marginal uniformity of the estimated copula distribution. We achieve this by learning a multivariate latent distribution with unspecified marginals but the desired dependency structure. By applying the probability integral transform, we can then obtain samples from the high-dimensional copula distribution without relying on parametric assumptions or the need to find a suitable tree structure. Experiments on synthetic and real data from finance, physics, and image generation demonstrate the performance of this approach.

Javeri, I. Y., Toutiaee, M., Arpinar, I. B., Miller, J. A., and Miller, T. W. (2021). “Improving Neural Networks for Time-Series Forecasting using Data Augmentation and AutoML.” In: *IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService)*. IEEE.

Statistical methods such as the Box-Jenkins method for time-series forecasting have been prominent since their development in 1970. Many researchers rely on such models as they can be efficiently estimated and also provide interpretability. However, advances in machine learning research indicate that neural networks can be powerful data modeling techniques, as they can provide higher accuracy for a plethora of learning problems and datasets. In the past, they have been tried on time-series forecasting as well, but their overall results have not been significantly better than the statistical models especially for intermediate length times-series data. Their modeling capacities are limited in cases where enough data may not be available to estimate the large number of parameters that these non-linear models require. This paper presents an easy to implement data augmentation method to significantly improve the performance of such networks. Our method, Augmented-Neural-Network, which involves using forecasts from statistical models, can help unlock the power of neural networks on intermediate length time-series and produces competitive results. It shows that data augmentation, when paired with Automated Machine Learning techniques such as Neural Architecture Search, can help to find the best neural architecture for a given time-series. Using the combination of these, demonstrates significant enhancement in the forecasting accuracy of three neural network-based models for a COVID-19 dataset, with a maximum improvement in forecasting accuracy by 21.41%, 24.29%, and 16.42%, respectively, over the neural networks that do not use augmented data.

Jericevich, I., McKechnie, M., and Gebbie, T. (2021). “Calibrating an adaptive Farmer-Joshi agent-based model for financial markets.” In: *arXiv e-Print*.

We replicate the contested calibration of the Farmer and Joshi agent based model of financial markets using a genetic algorithm and a Nelder-Mead with threshold accepting algorithm following Fabretti. The novelty of the Farmer-Joshi model is that the dynamics are driven by trade entry and exit thresholds alone. We recover the known claim that some important stylized facts observed in financial markets cannot be easily found under calibration – in particular those relating to the auto-correlations in the absolute values of the price fluctuations, and sufficient kurtosis. However, rather than concerns relating to the calibration method, what is novel here is that we extended the Farmer-Joshi model to include agent adaptation using an Brock and Hommes approach to strategy fitness based on trading strategy profitability. We call this an adaptive Farmer-Joshi model: the model allows trading agents to switch between strategies by favouring strategies that have been more profitable over some period of time determined by a free-parameter fixing the profit monitoring time-horizon. In the adaptive model we are able to calibrate and recover additional stylized facts, despite apparent degeneracy’s. This is achieved by combining the interactions of trade entry levels with trade strategy switching. We use this to argue that for low-frequency trading across days, as calibrated to daily sampled data, feed-backs can be accounted for by strategy die-out based on intermediate term profitability; we find that the average trade monitoring horizon is approximately two to three months (or 40 to 60 days) of trading.

Jiang, L., Dai, B., Wu, W., and Loy, C. C. (2021a). “Deceive D: Adaptive Pseudo Augmentation for GAN Training with Limited Data.” In: *arXiv e-Print*.

Generative adversarial networks (GANs) typically require ample data for training in order to synthesize high-fidelity images. Recent studies have shown that training GANs with limited data remains formidable due to



discriminator overfitting, the underlying cause that impedes the generator’s convergence. This paper introduces a novel strategy called Adaptive Pseudo Augmentation (APA) to encourage healthy competition between the generator and the discriminator. As an alternative method to existing approaches that rely on standard data augmentations or model regularization, APA alleviates overfitting by employing the generator itself to augment the real data distribution with generated images, which deceives the discriminator adaptively. Extensive experiments demonstrate the effectiveness of APA in improving synthesis quality in the low-data regime. We provide a theoretical analysis to examine the convergence and rationality of our new training strategy. APA is simple and effective. It can be added seamlessly to powerful contemporary GANs, such as StyleGAN2, with negligible computational cost.

Jiang, W., Song, S., Hou, L., and Zhao, H. (2021b). “A set of efficient methods to generate high-dimensional binary data with specified correlation structures.” In: *The American Statistician* 75(3), pp. 310-3221–37.

High dimensional correlated binary data arise in many areas, such as observed genetic variations in biomedical research. Data simulation can help researchers evaluate efficiency and explore properties of different computational and statistical methods. Also, some statistical methods, such as Monte-Carlo methods, rely on data simulation. Lunn and Davies (1998) proposed linear time complexity methods to generate correlated binary variables with three common correlation structures. However, it is infeasible to specify unequal probabilities in their methods. In this manuscript, we introduce several computationally efficient algorithms that generate high-dimensional binary data with specified correlation structures and unequal probabilities. Our algorithms have linear time complexity with respect to the dimension for three commonly studied correlation structures, namely exchangeable, decaying-product and K-dependent correlation structures. In addition, we extend our algorithms to generate binary data of specified non-negative correlation matrices satisfying the validity condition with quadratic time complexity. We provide an R package, CorBin, to implement our simulation methods. Compared to the existing packages for binary data generation, the time cost to generate a 100-dimensional binary vector with the common correlation structures and general correlation matrices can be reduced up to 10 to power 5 folds and 10 to power 3 folds, respectively, and the efficiency can be further improved with the increase of dimensions. The R package CorBin is available on CRAN at <https://cran.r-project.org/web/packages/CorBin/index.html>.

Jin, C. and Rinard, M. (2021). “Towards Context-Agnostic Learning Using Synthetic Data.” In: *arXiv e-Print*.

We propose a novel setting for learning, where the input domain is the image of a map defined on the product of two sets, one of which completely determines the labels. We derive a new risk bound for this setting that decomposes into a bias and an error term, and exhibits a surprisingly weak dependence on the true labels. Inspired by these results, we present an algorithm aimed at minimizing the bias term by exploiting the ability to sample from each set independently. We apply our setting to visual classification tasks, where our approach enables us to train classifiers on datasets that consist entirely of a single synthetic example of each class. On several standard benchmarks for real-world image classification, we achieve robust performance in the context-agnostic setting, with good generalization to real world domains, whereas training directly on real world data without our techniques yields classifiers that are brittle to perturbations of the background.

Johnson, M. C. and West, M. (2018). “Bayesian Predictive Synthesis: Forecast Calibration and Combination.” In: *arXiv e-Print*.

The combination of forecast densities, whether they result from a set of models, a group of consulted experts, or other sources, is becoming increasingly important in the fields of economics, policy and finance, among others. Requiring methodology that goes beyond standard Bayesian model uncertainty and model mixing - with its well-known limitations based on a clearly proscribed theoretical basis - multiple ‘density combination’ methods have been proposed. While some proposals have demonstrated empirical success, most apparently lack a core philosophical and theoretical foundation. Interesting recent examples generalize the common ‘linear opinion pool’ with flexible mixing weights that depend on the forecast variable itself - i.e., outcome-dependent mixing. Taking a foundational subjective Bayesian perspective, we show that such a density combination scheme is in fact justified as one example of Bayesian agent opinion analysis, or ‘predictive synthesis’. This logically coherent framework clearly delineates the underlying assumptions as well as the theoretical constraints and limitations of many combination ‘rules’, defining a broad class of Bayesian models for the general problem. A number of examples, including an application to a set of predictive densities in foreign exchange, provide illustrations.

Juan, A. A., Keenan, P., Martí, R., McGarraghy, S., Panadero, J., Carroll, P., and Oliva, D. (2022). “A review of the role of heuristics in stochastic optimisation: from metaheuristics to learnheuristics.” In: *Annals of Operations Research*.

In the context of simulation-based optimisation, this paper reviews recent work related to the role of metaheuristics, matheuristics (combinations of exact optimisation methods with metaheuristics), simheuristics (hybridisation of simulation with metaheuristics), biased-randomised heuristics for 'agile' optimisation via parallel computing, and learnheuristics (combination of statistical/machine learning with metaheuristics) to deal with NP-hard and large-scale optimisation problems in areas such as transport and logistics, manufacturing and production, smart cities, telecommunication networks, finance and insurance, sustainable energy consumption, health care, military and defence, e-marketing, or bioinformatics. The manuscript provides the main related concepts and updated references that illustrate the applications of these hybrid optimisation-simulation-learning approaches in solving rich and real-life challenges under dynamic and uncertainty scenarios. A numerical analysis is also included to illustrate the benefits that these approaches can offer across different application fields. Finally, this work concludes by highlighting open research lines on the combination of these methodologies to extend the concept of simulation-based optimisation.

- Kang, Y., Hyndman, R. J., and Li, F. (2020). “GRATIS: GeneRAtIng Time Series with diverse and controllable characteristics.” In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 13(4), pp. 354–376.

The explosion of time series data in recent years has brought a flourish of new time series analysis methods, for forecasting, clustering, classification and other tasks. The evaluation of these new methods requires either collecting or simulating a diverse set of time series benchmarking data to enable reliable comparisons against alternative approaches. We propose GeneRAtIng Time Series with diverse and controllable characteristics, named GRATIS, with the use of mixture autoregressive (MAR) models. We simulate sets of time series using MAR models and investigate the diversity and coverage of the generated time series in a time series feature space. By tuning the parameters of the MAR models, GRATIS is also able to efficiently generate new time series with controllable features. In general, as a costless surrogate to the traditional data collection approach, GRATIS can be used as an evaluation tool for tasks such as time series forecasting and classification. We illustrate the usefulness of our time series generation process through a time series forecasting application.

- Kang, Y., Spiliotis, E., Petropoulos, F., Athinoti, N., Li, F., and Assimakopoulos, V. (2021). “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.

- Kantos, C. and diBartolomeo, D. (2020). “How the pandemic taught us to turn smart beta into real alpha.” In: *Journal of Asset Management* 21(7), pp. 581–590.

The ongoing COVID-19 pandemic has strongly reminded equity investors that rare but extreme events occur from time to time. At the individual firm level, such events also impact the likelihood of bankruptcy, a feature that is not well represented in the traditional Capital Asset Pricing Model. This paper presents a functional form for equity asset pricing that is realistic, and reconciles the observed high equity risk premium with the observed lower than expected slope of the Security Market Line. Most importantly, we will demonstrate how including the potential for such large events changes traditional views of equity returns and the known factors that contribute to those returns. On the basis of empirical examination of a dataset stretching over 30 years without survivorship bias, we conclude that when the probabilities of rare extreme events are considered, strategies that focus on “alpha” (risk adjusted return) as defined in Jensen (J Finance 23(2):389-416, 1967) are structurally superior to “smart beta” strategies that seek to outperform a market index benchmark.

- Kapoor, A. and Shrivastava, U. (2014). “Extreme Values Theory and Return Level Analysis for Catastrophe Prediction.” In: *The Journal of Investing* 23(2), pp. 124–135.

As the trade and information flows in the financial markets have grown, they have become increasingly volatile and sensitive to extreme events. Markets are now characterized by more frequent extreme events. Generally, such extreme events in the market are not normally distributed and need to be modeled separately. Extreme value theory (EVT) is highly successful and is dedicated to modeling these catastrophic events. The question

now is which extreme values can be better estimated monthly, quarterly, or yearly. Hence, in this study, returns are divided into blocks of various sizes, and minimum return in these blocks is modeled using GEV distribution. Gumbel, Frechet, and Weibull type GEV distribution parameters for blocks of minimum monthly, quarterly, and yearly returns are estimated for NASDAQ returns of the last 30 years. The study also seeks to determine the accuracy of return level estimates obtained using GEV distribution. Return levels can be useful in characterizing bearish and bullish trends and predicting the same.

Karaś, M. and Serwatka, A. (2021). “Strong-hand conjecture: agent-based numerical simulation.” In: *The Journal of Investment Strategies*.

Following the example of the Kim-Markowitz model, this study adopts similar mechanisms of market operation to perform computer simulations based on agent modeling on the financial market, where shares of one company and a bank account are available (agents can invest and borrow money). The aim of the paper is to observe how the total portfolio changes in two groups of investors: so-called large and small investors. We examine the ratio of the total portfolio of large investors over time and hypothesize that, looking at two groups of investors who are equally rich at time  $t = 0$ , the group of large investors will gain more by the last day of our simulations. Our simulations show that large investors have a greater influence on stock prices and are stronger as a group, with a total portfolio greater than that of small investors on the last day of simulations, confirming the conjecture of strong hands.

Kegel, L., Hahmann, M., and Lehner, W. (2017). “Generating What-If Scenarios for Time Series Data.” In: *Proceedings of the 29th International Conference on Scientific and Statistical Database Management - '17*. New York, New York, USA: ACM Press, pp. 1–12.

Time series data has become a ubiquitous and important data source in many application domains. Most companies and organizations strongly rely on this data for critical tasks like decision-making, planning, predictions, and analytics in general. While all these tasks generally focus on actual data representing organization and business processes, it is also desirable to apply them to alternative scenarios in order to prepare for developments that diverge from expectations or assess the robustness of current strategies. When it comes to the construction of such what-if scenarios, existing tools either focus on scalar data or they address highly specific scenarios. In this work, we propose a generally applicable and easy-to-use method for the generation of what-if scenarios on time series data. Our approach extracts descriptive features of a data set and allows the construction of an alternate version by means of filtering and modification of these features.

Kegel, L., Hahmann, M., and Lehner, W. (2018). “Feature-based comparison and generation of time series.” In: *Proceedings of the 30th International Conference on Scientific and Statistical Database Management - '18*. New York, New York, USA: ACM Press, pp. 1–12.

For more than three decades, researchers have been developing generation methods for the weather, energy, and economic domain. These methods provide generated datasets for reasons like system evaluation and data availability. However, despite the variety of approaches, there is no comparative and cross-domain assessment of generation methods and their expressiveness. We present a similarity measure that analyzes generation methods regarding general time series features. By this means, users can compare generation methods and validate whether a generated dataset is considered similar to a given dataset. Moreover, we propose a feature-based generation method that evolves cross-domain time series datasets. This method outperforms other generation methods regarding the feature-based similarity.

Kemp, M. (2014). *Extreme Events: Robust Portfolio Construction in the Presence of Fat Tails*. Wiley. 334 pp.

Taking due account of extreme events when constructing portfolios of assets or liabilities is a key discipline for market professionals. Extreme events are a fact of life in how markets operate. In *Extreme Events: Robust Portfolio Construction in the Presence of Fat Tails*, leading expert Malcolm Kemp shows readers how to analyse market data to uncover fat-tailed behaviour, how to incorporate expert judgement in the handling of such information, and how to refine portfolio construction methodologies to make portfolios less vulnerable to extreme events or to benefit more from them. This is the only text that combines a comprehensive treatment of modern risk budgeting and portfolio construction techniques with the specific refinements needed for them to handle extreme events. It explains in a logical sequence what constitutes fat-tailed behaviour and why it arises, how we can analyse such behaviour, at aggregate, sector or instrument level, and how we can then take advantage of this analysis. Along the way, it provides a rigorous, comprehensive and clear development of traditional portfolio construction methodologies applicable if fat-tails are absent. It then explains how to refine these methodologies to accommodate real world behaviour. Throughout, the book highlights the importance of expert opinion, showing

that even the most data-centric portfolio construction approaches ultimately depend on practitioner assumptions about how the world might behave.

Kingma, D. P. and Welling, M. (2019). “An introduction to variational autoencoders.” In: *Foundations and Trends in Machine Learning* 12(4), pp. 307–392.

Variational autoencoders provide a principled framework for learning deep latent-variable models and corresponding inference models. In this work, we provide an introduction to variational autoencoders and some important extensions.

Koesdwiady, A., Khatib, A. E., and Kararay, F. (2018). “Methods to Improve Multi-Step Time Series Prediction.” In: *International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–8.

Multi-step time series prediction is known to suffer from increasing performance degradation the farther in the future the predictions are made. In this paper, we introduce two approaches to address this weakness in recursive and multioutput prediction models. In particular, we present a model that allows recursive prediction approaches to take into account the time-step index when making predictions. In addition, we propose a conditional generative adversarial network-based data augmentation model to improve prediction performance in multioutput models. We show on real-world time series datasets that the two methods improve on multi-step time series prediction in recursive and multi-output models, respectively.

Kondratyev, A. and Schwarz, C. (2020). “The Market Generator.” In: *Risk*.

We propose to use a special type of generative neural network - a Restricted Boltzmann Machine (RBM) - to build a powerful generator of synthetic market data that can replicate the probability distribution of the original market data. An RBM constructed with stochastic binary activation units in both the hidden and the visible layers (Bernoulli RBM) can learn complex dependence structures while avoiding overfitting. In this paper we consider an efficient data transformation and sampling approach that allows us to operate Bernoulli RBM on real-valued data sets and control the degree of autocorrelation and non-stationarity in the generated time series.

Kong, K., Li, G., Ding, M., Wu, Z., Zhu, C., Ghanem, B., Taylor, G., and Goldstein, T. (2022). “Robust Optimization as Data Augmentation for Large-scale Graphs.” In: *arXiv e-Print*.

Data augmentation helps neural networks generalize better by enlarging the training set, but it remains an open question how to effectively augment graph data to enhance the performance of GNNs (Graph Neural Networks). While most existing graph regularizers focus on manipulating graph topological structures by adding/removing edges, we offer a method to augment node features for better performance. We propose FLAG (Free Large-scale Adversarial Augmentation on Graphs), which iteratively augments node features with gradient-based adversarial perturbations during training. By making the model invariant to small fluctuations in input data, our method helps models generalize to out-of-distribution samples and boosts model performance at test time. FLAG is a general-purpose approach for graph data, which universally works in node classification, link prediction, and graph classification tasks. FLAG is also highly flexible and scalable, and is deployable with arbitrary GNN backbones and large-scale datasets. We demonstrate the efficacy and stability of our method through extensive experiments and ablation studies. We also provide intuitive observations for a deeper understanding of our method.

Koochali, A., Dengel, A., and Ahmed, S. (2020). “If You Like It, GAN It. Probabilistic Multivariate Times Series Forecast With GAN.” In: *arXiv e-Print*.

The contribution of this paper is two-fold. First, we present ProbCast - a novel probabilistic model for multivariate time-series forecasting. We employ a conditional GAN framework to train our model with adversarial training. Second, we propose a framework that lets us transform a deterministic model into a probabilistic one with improved performance. The motivation of the framework is to either transform existing highly accurate point forecast models to their probabilistic counterparts or to train GANs stably by selecting the architecture of GAN’s component carefully and efficiently. We conduct experiments over two publicly available datasets namely electricity consumption dataset and exchange-rate dataset. The results of the experiments demonstrate the remarkable performance of our model as well as the successful application of our proposed framework.

Kopeliovich, Y., Novosyolov, A., Satchkov, D., and Schachter, B. (2015). “Robust Risk Estimation and Hedging: A Reverse Stress Testing Approach.” In: *The Journal of Derivatives* 22(4), pp. 10–25.

The stress test has become an increasingly important risk assessment and management tool. But while it is easy to imagine a stress scenario and to estimate its impact on the firm’s financial condition, it is not so obvious how to select the most meaningful scenarios in the first place, either to get reasonable coverage of the space of stressful possibilities or even to focus on those that are most probable. In this article, the authors approach the problem from the reverse direction. They begin with a specified level of loss and pick the most likely scenario

that generates that loss. They then use principal components to construct a set of alternative scenarios that produce the same level of loss but in (maximally) different ways. This provides much greater insight into which sources of risk are the most important and the most stable across scenarios.

Koshiyama, A., Firoozye, N., and Treleaven, P. (2021). “Generative Adversarial Networks for Financial Trading Strategies Fine-Tuning and Combination.” In: *Quantitative Finance* 21(5), pp. 797–813.

Systematic trading strategies are algorithmic procedures that allocate assets aiming to optimize a certain performance criterion. To obtain an edge in a highly competitive environment, the analyst needs to properly fine-tune its strategy, or discover how to combine weak signals in novel alpha creating manners. Both aspects, namely fine-tuning and combination, have been extensively researched using several methods, but emerging techniques such as Generative Adversarial Networks can have an impact into such aspects. Therefore, our work proposes the use of Conditional Generative Adversarial Networks (cGANs) for trading strategies calibration and aggregation. To this purpose, we provide a full methodology on: (i) the training and selection of a cGAN for time series data; (ii) how each sample is used for strategies calibration; and (iii) how all generated samples can be used for ensemble modelling. To provide evidence that our approach is well grounded, we have designed an experiment with multiple trading strategies, encompassing 579 assets. We compared cGAN with an ensemble scheme and model validation methods, both suited for time series. Our results suggest that cGANs are a suitable alternative for strategies calibration and combination, providing outperformance when the traditional techniques fail to generate any alpha.

Kreiss, J.-P. and Paparoditis, E. (2011). “Bootstrap methods for dependent data: A review.” In: *Journal of the Korean Statistical Society* 40(4), pp. 357–378.

This paper gives a review on a variety of bootstrap methods for dependent data. The main focus is not on an exhaustive listing and description of bootstrap procedures but on general principles which should be taken into account when selecting a particular bootstrap procedure in order to approximate the (properly standardized) distribution of a statistic of interest. Questions are considered related to which dependence properties of the underlying data generating process asymptotically influence the distribution of the statistic of interest and which dependence properties (or even which process) a particular bootstrap method really mimics. For answering these questions we introduce the concept of a companion stochastic process. As statistics we consider generalized means, and integrated periodogram statistics (including ratio statistics) as well as nonparametric estimators.

Kritzman, M., Li, D., Qiu, G. (, and Turkington, D. (2021). “Portfolio Choice with Path-Dependent Scenarios.” In: *Financial Analysts Journal* 77(1), pp. 90–100.

Sophisticated investors rely on scenario analysis to select portfolios. We propose a new approach to scenario analysis that enables investors to consider sequential outcomes. We define scenarios not as average values but as paths for the economic variables. And we measure the likelihood of these paths on the basis of the statistical similarity of the paths to historical sequences. We also use a novel forecasting technique called “partial sample regression” to map economic outcomes onto asset class returns. This process allows investors to evaluate portfolios on the basis of the likelihood that the scenario will produce a certain pattern of returns over a specified investment horizon.

Kritzman, M. and Turkington, D. (2021). “History, Shocks and Drifts: A New Approach to Portfolio Formation.” In: *SSRN e-Print*.

Investors intuitively view future possibilities as a combination of historical outcomes, shocks that occur suddenly, and drifts that unfold gradually over several years. The authors show how to build portfolios based on such a view of the future. Their key innovation is to create a mixed-frequency return sample that properly balances short-term and long-term returns, and to form portfolios by considering all the returns of the sample instead of a statistical summary of them.

Kritzman, M. and Turkington, D. (2022). “History, Shocks, and Drifts: A New Approach to Portfolio Formation.” In: *The Journal of Portfolio Management* 48(2).

Investors intuitively view future possibilities as a combination of historical outcomes, shocks that occur suddenly, and drifts that unfold gradually over several years. The authors show how to build portfolios based on such a view of the future. Their key innovation is to create a mixed-frequency return sample that properly balances short-term and long-term returns and to form portfolios by considering all the returns of the sample instead of a statistical summary of them.

Kuchnik, M. and Smith, V. (2018). “Efficient Augmentation via Data Subsampling.” In: *arXiv e-Print*.

Data augmentation is commonly used to encode invariances in learning methods. However, this process is often performed in an inefficient manner, as artificial examples are created by applying a number of transformations



to all points in the training set. The resulting explosion of the dataset size can be an issue in terms of storage and training costs, as well as in selecting and tuning the optimal set of transformations to apply. In this work, we demonstrate that it is possible to significantly reduce the number of data points included in data augmentation while realizing the same accuracy and invariance benefits of augmenting the entire dataset. We propose a novel set of subsampling policies, based on model influence and loss, that can achieve a 90% reduction in augmentation set size while maintaining the accuracy gains of standard data augmentation.

Lam, H. and Li, F. (2020). “Parametric Scenario Optimization under Limited Data: A Distributionally Robust Optimization View.” In: *arXiv e-Print*.

We consider optimization problems with uncertain constraints that need to be satisfied probabilistically. When data are available, a common method to obtain feasible solutions for such problems is to impose sampled constraints, following the so-called scenario optimization approach. However, when the data size is small, the sampled constraints may not statistically support a feasibility guarantee on the obtained solution. This paper studies how to leverage parametric information and the power of Monte Carlo simulation to obtain feasible solutions for small-data situations. Our approach makes use of a distributionally robust optimization (DRO) formulation that translates the data size requirement into a Monte Carlo sample size requirement drawn from what we call a generating distribution. We show that, while the optimal choice of this generating distribution is the one eliciting the data or the baseline distribution in a nonparametric divergence-based DRO, it is not necessarily so in the parametric case. Correspondingly, we develop procedures to obtain generating distributions that improve upon these basic choices. We support our findings with several numerical examples.

Laskin, M., Lee, K., Stooke, A., Pinto, L., Abbeel, P., and Srinivas, A. (2020). “Reinforcement Learning with Augmented Data.” In: *arXiv e-Print*.

Learning from visual observations is a fundamental yet challenging problem in reinforcement learning (RL). Although algorithmic advancements combined with convolutional neural networks have proved to be a recipe for success, current methods are still lacking on two fronts: (a) sample efficiency of learning and (b) generalization to new environments. To this end, we present RAD: Reinforcement Learning with Augmented Data, a simple plug-and-play module that can enhance any RL algorithm. We show that data augmentations such as random crop, color jitter, patch cutout, and random convolutions can enable simple RL algorithms to match and even outperform complex state-of-the-art methods across common benchmarks in terms of data-efficiency, generalization, and wall-clock speed. We find that data diversity alone can make agents focus on meaningful information from high-dimensional observations without any changes to the reinforcement learning method. On the DeepMind Control Suite, we show that RAD is state-of-the-art in terms of data-efficiency and performance across 15 environments. We further demonstrate that RAD can significantly improve the test-time generalization on several OpenAI ProcGen benchmarks. Finally, our customized data augmentation modules enable faster wall-clock speed compared to competing RL techniques. Our RAD module and training code are available at <https://github.com/MishaLaskin/rad>.

Lawrence, A. R., Kaiser, M., Sampaio, R., and Sipos, M. (2021). “Data Generating Process to Evaluate Causal Discovery Techniques for Time Series Data.” In: *arXiv e-Print*.

Going beyond correlations, the understanding and identification of causal relationships in observational time series, an important subfield of Causal Discovery, poses a major challenge. The lack of access to a well-defined ground truth for real-world data creates the need to rely on synthetic data for the evaluation of these methods. Existing benchmarks are limited in their scope, as they either are restricted to a “static” selection of data sets, or do not allow for a granular assessment of the methods’ performance when commonly made assumptions are violated. We propose a flexible and simple to use framework for generating time series data, which is aimed at developing, evaluating, and benchmarking time series causal discovery methods. In particular, the framework can be used to fine tune novel methods on vast amounts of data, without “overfitting” them to a benchmark, but rather so they perform well in real-world use cases. Using our framework, we evaluate prominent time series causal discovery methods and demonstrate a notable degradation in performance when their assumptions are invalidated and their sensitivity to choice of hyperparameters. Finally, we propose future research directions and how our framework can support both researchers and practitioners.

Ledermann, D. and Alexander, C. (2012). “Further properties of random orthogonal matrix simulation.” In: *Mathematics and Computers in Simulation* 83, pp. 56–79.

Random orthogonal matrix (ROM) simulation is a very fast procedure for generating multivariate random samples that always have exactly the same mean, covariance and Mardia multivariate skewness and kurtosis. This paper investigates how the properties of parametric, data-specific and deterministic ROM simulations



are influenced by the choice of orthogonal matrix. Specifically, we consider how cyclic and general permutation matrices alter their time-series properties, and how three classes of rotation matrices - upper Hessenberg, Cayley, and exponential - influence both the unconditional moments of the marginal distributions and the behaviour of skewness when samples are concatenated. We also perform an experiment which demonstrates that parametric ROM simulation can be hundreds of times faster than equivalent Monte Carlo simulation.

Ledermann, W., Alexander, C., and Ledermann, D. (2011). "Random orthogonal matrix simulation." In: *Linear Algebra and its Applications* 434, pp. 1444–1467.

This paper introduces a method for simulating multivariate samples that have exact means, covariances, skewness and kurtosis. We introduce a new class of rectangular orthogonal matrices which is fundamental to the methodology and we call these matrices L matrices. They may be deterministic, parametric or data specific in nature. The target moments determine the L matrix then infinitely many random samples with the exact same moments may be generated by multiplying the L matrix by arbitrary random orthogonal matrices. This methodology is thus termed ROM simulation. Considering certain elementary types of random orthogonal matrices we demonstrate that they generate samples with different characteristics. ROM simulation has applications to many problems that are resolved using standard Monte Carlo methods. But no parametric assumptions are required (unless parametric L matrices are used) so there is no sampling error caused by the discrete approximation of a continuous distribution, which is a major source of error in standard Monte Carlo simulations. For illustration, we apply ROM simulation to determine the value-at-risk of a stock portfolio.

Ledoit, O. and Wolf, M. (2021). "Shrinkage estimation of large covariance matrices: Keep it simple, statistician?" In: *Journal of Multivariate Analysis* 186, p. 104796.

Under rotation-equivariant decision theory, sample covariance matrix eigenvalues can be optimally shrunk by recombining sample eigenvectors with a (potentially nonlinear) function of the unobservable population covariance matrix. The optimal shape of this function reflects the loss/risk that is to be minimized. We solve the problem of optimal covariance matrix estimation under a variety of loss functions motivated by statistical precedent, probability theory, and differential geometry. A key ingredient of our nonlinear shrinkage methodology is a new estimator of the angle between sample and population eigenvectors, without making strong assumptions on the population eigenvalues. We also introduce a broad family of covariance matrix estimators that can handle all regular functional transformations of the population covariance matrix under large-dimensional asymptotics. In addition, we compare via Monte Carlo simulations our methodology to two simpler ones from the literature, linear shrinkage and shrinkage based on the spiked covariance model.

Lee, Y.-H., Hsieh, M.-H., Kuo, W., and Tsai, C. J. (2021). "How can an economic scenario generation model cope with abrupt changes in financial markets?" In: *China Finance Review International* 11(3), pp. 372–405.

Purpose: It is quite possible that financial institutions including life insurance companies would encounter turbulent situations such as the COVID-19 pandemic before policies mature. Constructing models that can generate scenarios for major assets to cover abrupt changes in financial markets is thus essential for the financial institution's risk management.

Design/methodology/approach: The key issues in such modeling include how to manage the large number of risk factors involved, how to model the dynamics of chosen or derived factors and how to incorporate relations among these factors. The authors propose the orthogonal ARMA-GARCH (autoregressive moving-average-generalized autoregressive conditional heteroskedasticity) approach to tackle these issues. The constructed economic scenario generation (ESG) models pass the backtests covering the period from the beginning of 2018 to the end of May 2020, which includes the turbulent situations caused by COVID-19.

Findings: The backtesting covering the turbulent period of COVID-19, along with fan charts and comparisons on simulated and historical statistics, validates our approach.

Originality/value: This paper is the first one that attempts to generate complex long-term economic scenarios for a large-scale portfolio from its large dimensional covariance matrix estimated by the orthogonal ARMA-GARCH model.

Lee, M. I. (2013). "Stress testing Monte Carlo assumptions." In: *SSRN e-Print*.

Despite the evidence that returns are fat-tailed and that expected returns vary through time, most Monte Carlo simulations assume returns are independent and identically normally distributed. This study incorporates these return patterns in retirement simulations to illustrate how common assumptions about returns impact the output of Monte Carlo simulations.

Leon, C. and Reveiz, A. (2012). "Monte Carlo Simulation of Long-Term Dependent Processes: A Primer." In: *Wilmott* 2012(60), pp. 48–57.

This article briefly describes the Cholesky method for simulating Geometric Brownian Motion processes with long-term dependence, also referred as Fractional Geometric Brownian Motion. This choice results from its parsimony, simplicity and documented theoretical advantages.

- 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.

- Lezmi, E., Malongo, H., Roncalli, T., and Sobotka, R. (2019). “Portfolio Allocation with Skewness Risk: A Practical Guide.” In: *SSRN e-Print*.

In this article, we show how to take into account skewness risk in portfolio allocation. Until recently, this issue has been seen as a purely statistical problem, since skewness corresponds to the third statistical moment of a probability distribution. However, in finance, the concept of skewness is more related to extreme events that produce portfolio losses. More precisely, the skewness measures the outcome resulting from bad times and adverse scenarios in financial markets. Based on this interpretation of the skewness risk, we focus on two approaches that are closely connected. The first one is based on the Gaussian mixture model with two regimes: a normal regime and a turbulent regime. The second approach directly incorporates a stress scenario using jump-diffusion modeling. This second approach can be seen as a special case of the first approach. However, it has the advantage of being clearer and more in line with the experience of professionals in financial markets: skewness is due to negative jumps in asset prices. After presenting the mathematical framework, we analyze an investment portfolio that mixes risk premia, more specifically risk parity, momentum and carry strategies. We show that traditional portfolio management based on the volatility risk measure is biased and corresponds to a short-sighted approach to bad times. We then propose to replace the volatility risk measure by a skewness risk measure, which is calculated as an expected shortfall that incorporates a stress scenario. We conclude that constant-mix portfolios may be better adapted than actively managed portfolios, when the investment universe is composed of negatively skewed financial assets.

- Lezmi, E., Roche, J., Roncalli, T., and Xu, J. (2020). “Improving the Robustness of Trading Strategy Backtesting with Boltzmann Machines and Generative Adversarial Networks.” In: *SSRN e-Print*.

This article explores the use of machine learning models to build a market generator. The underlying idea is to simulate artificial multi-dimensional financial time series, whose statistical properties are the same as those observed in the financial markets. In particular, these synthetic data must preserve the probability distribution of asset returns, the stochastic dependence between the different assets and the autocorrelation across time. The article proposes then a new approach for estimating the probability distribution of backtest statistics. The final objective is to develop a framework for improving the risk management of quantitative investment strategies, in particular in the space of smart beta, factor investing and alternative risk premia.

- Leznik, M., Michalsky, P., Willis, P., Schanzel, B., Ostberg, P.-O., and Domaschka, J. (2021). “Multivariate Time Series Synthesis Using Generative Adversarial Networks.” In: *Proceedings of the ACM/SPEC International Conference on Performance Engineering*. ACM.

Collection and analysis of distributed (cloud) computing workloads allows for a deeper understanding of user and system behavior and is necessary for efficient operation of infrastructures and applications. The availability of such workload data is however often limited as most cloud infrastructures are commercially operated and monitoring data is considered proprietary or falls under GDPR regulations. This work investigates the generation of synthetic workloads using Generative Adversarial Networks and addresses a current need for more data and better tools for workload generation. Resource utilization measurements such as the utilization rates of Content Delivery Network (CDN) caches are generated and a comparative evaluation pipeline using descriptive statistics and time-series analysis is developed to assess the statistical similarity of generated and measured workloads. We use CDN data open sourced by us in a data generation pipeline as well as back-end ISP workload data to

demonstrate the multivariate synthesis capability of our approach. The work contributes a generation method for multivariate time series workload generation that can provide arbitrary amounts of statistically similar data sets based on small subsets of real data. The presented technique shows promising results, in particular for heterogeneous workloads not too irregular in temporal behavior.

- Li, B., Luo, S., Qin, X., and Pan, L. (2021). “Improving GAN with inverse cumulative distribution function for tabular data synthesis.” In: *Neurocomputing* 456, pp. 373–383.

Designing a generative model to synthesize realistic tabular data is of great significance in data science. Existing tabular data generative models have difficulty in handling complicated and diverse marginal distribution types due to the gradient vanishing problem, and these models pay little attention to the correlation between attributes. We propose a method that improves the generative adversarial network (GAN) with inverse cumulative distribution function for tabular data synthesis. This method first transforms continuous columns into uniform distribution data by using the cumulative distribution function, which can alleviate the gradient vanishing problem in model training. Then the method trains GAN with the transformed data, where the discriminator with label reconstruction function is presented to model the correlation among attributes accurately by introducing an auxiliary supervised task to help the correlations extraction. After that, we train a neural network for each continuous column to perform the inverse transformation of generated data into the target distribution, thereby the synthetic data is obtained. Experiments on simulated and real-world datasets show that our method compares favorably against the state-of-the-art methods in modeling tabular data.

- Li, G., Huang, L., Yang, J., and Zhang, W. (2022). “A Synthetic Regression Model for Large Portfolio Allocation.” In: *Journal of Business & Economic Statistics*, pp. 1–13.

Portfolio allocation is an important topic in financial data analysis. In this article, based on the mean-variance optimization principle, we propose a synthetic regression model for construction of portfolio allocation, and an easy to implement approach to generate the synthetic sample for the model. Compared with the regression approach in existing literature for portfolio allocation, the proposed method of generating the synthetic sample provides more accurate approximation for the synthetic response variable when the number of assets under consideration is large. Due to the embedded leave-one-out idea, the synthetic sample generated by the proposed method has weaker within sample correlation, which makes the resulting portfolio allocation more close to the optimal one. This intuitive conclusion is theoretically confirmed to be true by the asymptotic properties established in this article. We have also conducted intensive simulation studies in this article to compare the proposed method with the existing ones, and found the proposed method works better. Finally, we apply the proposed method to real datasets. The yielded returns look very encouraging.

- Li, Z., Zhao, Y., and Fu, J. (2020). “SynC: A Copula based Framework for Generating Synthetic Data from Aggregated Sources.” In: *International Conference on Data Mining Workshops (ICDMW)*. IEEE.

A synthetic dataset is a data object that is generated programmatically, and it may be valuable to creating a single dataset from multiple sources when direct collection is difficult or costly. Although it is a fundamental step for many data science tasks, an efficient and standard framework is absent. In this paper, we study a specific synthetic data generation task called downscaling, a procedure to infer high-resolution, harder-to-collect information (e.g., individual level records) from many low-resolution, easy-to-collect sources, and propose a multi-stage framework called SynC (Synthetic Data Generation via Gaussian Copula). For given low-resolution datasets, the central idea of SynC is to fit Gaussian copula models to each of the low-resolution datasets in order to correctly capture dependencies and marginal distributions, and then sample from the fitted models to obtain the desired high-resolution subsets. Predictive models are then used to merge sampled subsets into one, and finally, sampled datasets are scaled according to low-resolution marginal constraints.

We make four key contributions in this work:

- 1) propose a novel framework for generating individual level data from aggregated data sources by combining state-of-the-art machine learning and statistical techniques,
- 2) perform simulation studies to validate SynC’s performance as a synthetic data generation algorithm,
- 3) demonstrate its value as a feature engineering tool, as well as an alternative to data collection in situations where gathering is difficult through two real-world datasets,
- 4) release an easy-to-use framework implementation (see <https://github.com/winston11/SynC>) for reproducibility and scalability at the production level that easily incorporates new data.

We have proposed a multi-stage framework called SynC (Synthetic Data Generation via Gaussian Copula).

Lim, S. H., Erichson, N. B., Utrera, F., Xu, W., and Mahoney, M. W. (2021). “Noisy Feature Mixup.” In: *arXiv e-Print*.

We introduce Noisy Feature Mixup (NFM), an inexpensive yet effective method for data augmentation that combines the best of interpolation based training and noise injection schemes. Rather than training with convex combinations of pairs of examples and their labels, we use noise-perturbed convex combinations of pairs of data points in both input and feature space. This method includes mixup and manifold mixup as special cases, but it has additional advantages, including better smoothing of decision boundaries and enabling improved model robustness. We provide theory to understand this as well as the implicit regularization effects of NFM. Our theory is supported by empirical results, demonstrating the advantage of NFM, as compared to mixup and manifold mixup. We show that residual networks and vision transformers trained with NFM have favorable trade-offs between predictive accuracy on clean data and robustness with respect to various types of data perturbation across a range of computer vision benchmark datasets.

Lin, Z., Jain, A., Wang, C., Fanti, G., and Sekar, V. (2019). “Generating High-fidelity, Synthetic Time Series Datasets with DoppelGANger.” In: *arXiv e-Print*.

Limited data access is a substantial barrier to data-driven networking research and development. Although many organizations are motivated to share data, privacy concerns often prevent the sharing of proprietary data, including between teams in the same organization and with outside stakeholders (e.g., researchers, vendors). Many researchers have therefore proposed synthetic data models, most of which have not gained traction because of their narrow scope. In this work, we present DoppelGANger, a synthetic data generation framework based on generative adversarial networks (GANs). DoppelGANger is designed to work on time series datasets with both continuous features (e.g. traffic measurements) and discrete ones (e.g., protocol name). Modeling time series and mixed-type data is known to be difficult; DoppelGANger circumvents these problems through a new conditional architecture that isolates the generation of metadata from time series, but uses metadata to strongly influence time series generation. We demonstrate the efficacy of DoppelGANger on three real-world datasets. We show that DoppelGANger achieves up to 43% better fidelity than baseline models, and captures structural properties of data that baseline methods are unable to learn. Additionally, it gives data holders an easy mechanism for protecting attributes of their data without substantial loss of data utility.

Linardi, M., Zhu, Y., Palpanas, T., and Keogh, E. (2018). “VALMOD: A suite for easy and exact detection of variable length motifs in data series.” In: *Proceedings of the 2018 International Conference on Management of Data - SIGMOD '18*. New York, New York, USA: ACM Press, pp. 1757–1760.

Data series motif discovery represents one of the most useful primitives for data series mining, with applications to many domains, such as robotics, entomology, seismology, medicine, and climatology, and others. The state-of-the-art motif discovery tools still require the user to provide the motif length. Yet, in several cases, the choice of motif length is critical for their detection. Unfortunately, the obvious brute-force solution, which tests all lengths within a given range, is computationally untenable, and does not provide any support for ranking motifs at different resolutions (i.e., lengths). We demonstrate VALMOD, our scalable motif discovery algorithm that efficiently finds all motifs in a given range of lengths, and outputs a length-invariant ranking of motifs. Furthermore, we support the analysis process by means of a newly proposed meta-data structure that helps the user to select the most promising pattern length. This demo aims at illustrating in detail the steps of the proposed approach, showcasing how our algorithm and corresponding graphical insights enable users to efficiently identify the correct motifs.

Lindner, D., Shah, R., Abbeel, P., and Dragan, A. (2021). “Learning What To Do by Simulating the Past.” In: *arXiv e-Print*.

Since reward functions are hard to specify, recent work has focused on learning policies from human feedback. However, such approaches are impeded by the expense of acquiring such feedback. Recent work proposed that agents have access to a source of information that is effectively free: in any environment that humans have acted in, the state will already be optimized for human preferences, and thus an agent can extract information about what humans want from the state. Such learning is possible in principle, but requires simulating all possible past trajectories that could have led to the observed state. This is feasible in gridworlds, but how do we scale it to complex tasks? In this work, we show that by combining a learned feature encoder with learned inverse models, we can enable agents to simulate human actions backwards in time to infer what they must have done. The resulting algorithm is able to reproduce a specific skill in MuJoCo environments given a single state sampled from the optimal policy for that skill.

Little, M. A. and Badawy, R. (2020). “Causal bootstrapping.” In: *arXiv e-Print*.

To draw scientifically meaningful conclusions and build reliable models of quantitative phenomena, cause and effect must be taken into consideration (either implicitly or explicitly). This is particularly challenging when the measurements are not from controlled experimental (interventional) settings, since cause and effect can be obscured by spurious, indirect influences. Modern predictive techniques from machine learning are capable of capturing high-dimensional, nonlinear relationships between variables while relying on few parametric or probabilistic model assumptions. However, since these techniques are associational, applied to observational data they are prone to picking up spurious influences from non-experimental (observational) data, making their predictions unreliable. Techniques from causal inference, such as probabilistic causal diagrams and do-calculus, provide powerful (nonparametric) tools for drawing causal inferences from such observational data. However, these techniques are often incompatible with modern, nonparametric machine learning algorithms since they typically require explicit probabilistic models. Here, we develop causal bootstrapping for augmenting classical nonparametric bootstrap resampling with information on the causal relationship between variables. This makes it possible to resample observational data such that, if it is possible to identify an interventional relationship from that data, new data representing that relationship can be simulated from the original observational data. In this way, we can use modern machine learning algorithms unaltered to make statistically powerful, yet causally-robust, predictions. We develop several causal bootstrapping algorithms for drawing interventional inferences from observational data, for classification and regression problems, and demonstrate, using synthetic and real-world examples, the value of this approach.

Liu, B., Zhang, Z., and Cui, R. (2020). “Efficient Time Series Augmentation Methods.” In: *13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE.

With the emergence of deep learning and the improvement of computing ability, artificial intelligence has also made great progress. In the case of sufficient data, deep learning models can achieve very good results. However, in some tasks with scarce resources, the results of these models are somewhat unsatisfactory. In this regard, researchers have proposed some data augmentation technologies in the domain of computer vision and natural language processing. However, there is still a lack of effective methods for data augmentation applied to time series data. In this paper, we propose several methods to augment time series data, which are called AddNoise, Permutation, Scaling, Warping. Then, we verify these augmentation methods by two deep learning models, Fully Convolutional Neural Network (FCN) and ResNet, on real time series datasets. The experimental results show that these methods are useful to train the above two models in time series classification task.

Lu, L. and Ghosh, S. (2021). “Nonparametric estimation of multivariate copula using empirical bayes method.” In: *arXiv e-Print*.

In the field of finance, insurance, and system reliability, etc., it is often of interest to measure the dependence among variables by modeling a multivariate distribution using a copula. The copula models with parametric assumptions are easy to estimate but can be highly biased when such assumptions are false, while the empirical copulas are non-smooth and often not genuine copula making the inference about dependence challenging in practice. As a compromise, the empirical Bernstein copula provides a smooth estimator but the estimation of tuning parameters remains elusive. In this paper, by using the so-called empirical checkerboard copula we build a hierarchical empirical Bayes model that enables the estimation of a smooth copula function for arbitrary dimensions. The proposed estimator based on the multivariate Bernstein polynomials is itself a genuine copula and the selection of its dimension-varying degrees is data-dependent. We also show that the proposed copula estimator provides a more accurate estimate of several multivariate dependence measures which can be obtained in closed form. We investigate the asymptotic and finite-sample performance of the proposed estimator and compare it with some nonparametric estimators through simulation studies. An application to portfolio risk management is presented along with a quantification of estimation uncertainty.

Lux, T. (2021). “Can heterogeneous agent models explain the alleged mispricing of the S&P 500?” In: *Quantitative Finance* 21(9), pp. 1413–1433.

Models with heterogeneous agents go some way in explaining the bi-modality of the distortion between the S&P 500 and its ex-post rational fundamental value.

Ma, J. (2021). “copent: Estimating Copula Entropy and Transfer Entropy in R.” In: *arXiv e-Print*.

Statistical independence and conditional independence are two fundamental concepts in statistics and machine learning. Copula Entropy is a mathematical concept defined by Ma and Sun for multivariate statistical independence measuring and testing, and also proved to be closely related to conditional independence (or transfer entropy). As the unified framework for measuring both independence and causality, CE has been applied to solve several related statistical or machine learning problems, including association discovery, structure learning,



variable selection, and causal discovery. The nonparametric methods for estimating copula entropy and transfer entropy were also proposed previously. This paper introduces `copent`, the R package which implements these proposed methods for estimating copula entropy and transfer entropy. The implementation detail of the package is introduced. Three examples with simulated data and real-world data on variable selection and causal discovery are also presented to demonstrate the usage of this package. The examples on variable selection and causal discovery show the strong ability of `copent` on testing (conditional) independence compared with the related packages. The `copent` package is available on the Comprehensive R Archive Network (CRAN) and also on GitHub at <https://github.com/majianthu/copent>.

- Mammen, E. and Nandi, S. (2012). “Bootstrap and Resampling.” In: *Handbook of Computational Statistics*. Ed. by J. E. Gentle, W. K. Hardle, and Y. Mori. Springer Handbooks of Computational Statistics. Springer Berlin Heidelberg, pp. 499–527.

The bootstrap is by now a standard method in modern statistics. Its roots go back to a lot of resampling ideas that were around in the seventies. The seminal work of Efron synthesized some of the earlier resampling ideas and established a new framework for simulation based statistical analysis. The idea of the bootstrap is to develop a setup to generate more (pseudo) data using the information of the original data. True underlying sample properties are reproduced as closely as possible and unknown model characteristics are replaced by sample estimates.

- Mannix, R. and Cesa, M. (2021). “‘Signatures’ promise quants a tool for all jobs.” In: *Risk (Quant Investing)*.

A technique known as signatures could reshape how quants describe and compare patterns in financial data. Alongside academics, practitioners at firms such as Citi, JP Morgan, Standard Chartered and UBS are among those taking an interest in the field. Signatures could find uses in pricing exotic options, calibrating models, mining for alpha signals or reducing the storage demands of high-frequency data. The approach can be computationally demanding, although mechanisms are being developed to address that problem. “The convergence between the maths, the availability of the tools to calculate signatures, and a growing level of understanding makes this a methodology to watch,” says a leading quant.

- Mariani, G., Zhu, Y., Li, J., Scheidegger, F., Istrate, R., Bekas, C., and Malossi, A. C. I. (2019). “PAGAN: Portfolio Analysis with Generative Adversarial Networks.” In: *arXiv e-Print*.

Since decades, the data science community tries to propose prediction models of financial time series. Yet, driven by the rapid development of information technology and machine intelligence, the velocity of today’s information leads to high market efficiency. Sound financial theories demonstrate that in an efficient marketplace all information available today, including expectations on future events, are represented in today prices whereas future price trend is driven by the uncertainty. This jeopardizes the efforts put in designing prediction models. To deal with the unpredictability of financial systems, today’s portfolio management is largely based on the Markowitz framework which puts more emphasis in the analysis of the market uncertainty and less in the price prediction. The limitation of the Markowitz framework stands in taking very strong ideal assumptions about future returns probability distribution. To address this situation we propose PAGAN, a pioneering methodology based on deep generative models. The goal is modeling the market uncertainty that ultimately is the main factor driving future trends. The generative model learns the joint probability distribution of price trends for a set of financial assets to match the probability distribution of the real market. Once the model is trained, a portfolio is optimized by deciding the best diversification to minimize the risk and maximize the expected returns observed over the execution of several simulations. Applying the model for analyzing possible futures, is as simple as executing a Monte Carlo simulation, a technique very familiar to finance experts. The experimental results on different portfolios representing different geopolitical areas and industrial segments constructed using real-world public data sets demonstrate promising results.

- Marti, G. (2020a). “CORRGAN: sampling realistic financial correlation matrices using generative adversarial networks.” In: *ICASSP - IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 8459–8463.

We propose a novel approach for sampling realistic financial correlation matrices. This approach is based on generative adversarial networks. Experiments demonstrate that generative adversarial networks are able to recover most of the known stylized facts about empirical correlation matrices estimated on asset returns. This is the first time such results are documented in the literature. Practical financial applications range from trading strategies enhancement to risk and portfolio stress testing. Such generative models can also help ground empirical finance deeper into science by allowing for falsifiability of statements and more objective comparison of empirical methods.

- Marti, G. (2020b). “Generating Realistic Synthetic Data in Finance: Applications of GANs.” In: *IHS Markit Webinar*. Talk at IHS Markit Webinar (15 October 2020) on the potential Applications of GANs in Finance. These models could be useful for quants and their managers to avoid over-fitting, portfolio and risk managers for proper capital and risk allocation, cloud computing servicing willing to work with banks and other sensitive data rich organizations, auditors and regulators to detect frauds, and data vendors (such as IHS Markit) to bring new products to market and iterate quickly with clients.
- Marti, G., Andler, S., Nielsen, F., and Donnat, P. (2017). “Exploring and measuring non-linear correlations: Copulas, Lightspeed Transportation and Clustering.” In: *Proceedings of Machine Learning Research* 55, pp. 59–69. We propose a methodology to explore and measure the pairwise correlations that exist between variables in a dataset. The methodology leverages copulas for encoding dependence between two variables, state-of-the-art optimal transport for providing a relevant geometry to the copulas, and clustering for summarizing the main dependence patterns found between the variables. Some of the clusters centers can be used to parameterize a novel dependence coefficient which can target or forget specific dependence patterns. Finally, we illustrate and benchmark the methodology on several datasets. Code and numerical experiments are available online at <https://www.datagrapple.com/Tech> for reproducible research.
- Marti, G., Goubet, V., and Nielsen, F. (2021a). “cCorrGAN: Conditional Correlation GAN for Learning Empirical Conditional Distributions in the Elliptope.” In: *International Conference on Geometric Science of Information*. Springer International Publishing, pp. 613–620. We propose a methodology to approximate conditional distributions in the elliptope of correlation matrices based on conditional generative adversarial networks. We illustrate the methodology with an application from quantitative finance: Monte Carlo simulations of correlated returns to compare risk-based portfolio construction methods. Finally, we discuss about current limitations and advocate for further exploration of the elliptope geometry to improve results.
- Marti, G., Nielsen, F., Bihkowski, M., and Donnat, P. (2021b). “A review of two decades of correlations, hierarchies, networks and clustering in financial markets.” In: *Progress in Information Geometry*, pp. 245–274. We review the state of the art of clustering financial time series and the study of their correlations alongside other interaction networks. The aim of the review is to gather in one place the relevant material from different fields, e.g. machine learning, information geometry, econophysics, statistical physics, econometrics, behavioral finance. We hope it will help researchers to use more effectively this alternative modeling of the financial time series. Decision makers and quantitative researchers may also be able to leverage its insights. Finally, we also hope that this review will form the basis of an open toolbox to study correlations, hierarchies, networks and clustering in financial markets.
- Medina, M. A., Davis, R. A., and Samorodnitsky, G. (2021). “Spectral learning of multivariate extremes.” In: *arXiv e-Print*. We propose a spectral clustering algorithm for analyzing the dependence structure of multivariate extremes. More specifically, we focus on the asymptotic dependence of multivariate extremes characterized by the angular or spectral measure in extreme value theory. Our work studies the theoretical performance of spectral clustering based on a random  $k$ -nearest neighbor graph constructed from an extremal sample, i.e., the angular part of random vectors for which the radius exceeds a large threshold. In particular, we derive the asymptotic distribution of extremes arising from a linear factor model and prove that, under certain conditions, spectral clustering can consistently identify the clusters of extremes arising in this model. Leveraging this result we propose a simple consistent estimation strategy for learning the angular measure. Our theoretical findings are complemented with numerical experiments illustrating the finite sample performance of our methods.
- Meucci, A. (2008). “Fully Flexible Views: Theory and Practice.” In: *Risk* 21(10), pp. 97–102. We propose a unified methodology to input non-linear views from any number of users in fully general non-normal markets, and perform, among others, stress-testing, scenario analysis, and ranking allocation. We walk the reader through the theory and we detail an extremely efficient algorithm to easily implement this methodology under fully general assumptions. As it turns out, no repricing is ever necessary, hence the methodology can be readily applied to books with complex derivatives. We also present an analytical solution, useful for benchmarking, which per se generalizes notable previous results. Code illustrating this methodology in practice is available through author’s homepage.
- Meucci, A. (2012). *A Fully Integrated Liquidity and Market Risk Model*. Tech. rep. EDHEC. We introduce a new framework to integrate liquidity risk, funding risk and market risk, which goes beyond the simple bid-ask spread overlay to a VaR number. In our approach, we overlay a whole distribution of liquidity

uncertainty to each future market-risk scenario. Then we allow for the liquidity uncertainty to vary scenario by scenario, depending on what liquidation policy or funding policy is implemented in that scenario. The result is one easy-to-interpret and easy-to-implement formula for the total liquidity-plus-market-risk P&L distribution. Using this formula we can stress-test different market risk P&L distributions and different scenario-dependent liquidation policies and funding policies; compute total risk and decompose it into a novel liquidity-plus-market risk formula; and define a liquidity score as a monetary measure of portfolio liquidity. Our approach relies on three pillars: first, the literature on optimal execution, to model liquidity risk as a function of the actual trading involved; second, an analytical conditional convolution, to blend market risk and liquidity/funding risk; third the Fully Flexible Probabilities framework, to model and stress-test market risk even in highly non-normal portfolios with complex derivatives. Our approach can be implemented efficiently with portfolios of thousand of securities. The code for the case study is available for download.

Meucci, A., Ardia, D., and Keel, S. (2012). “Fully Flexible Extreme Views.” In: *Risk* 14(2), pp. 39–49.

We extend the Fully Flexible Views generalization of the Black-Litterman approach to effectively handle extreme views on the tails of a distribution. First, we provide a recursive algorithm to process views on the conditional value at risk, which cannot be handled directly by the original implementation of Fully Flexible Views. Second, we represent both the prior and the posterior distribution on a grid, instead of by means of Monte Carlo scenarios: this way it becomes possible to cover parsimoniously even the far tails of the underlying distribution. Documented code is available for download.

Meyer, D. and Nagler, T. (2021). “Synthia: multidimensional synthetic data generation in Python.” In: *Journal of Open Source Software* 6(65), p. 2863.

Synthetic data - artificially generated data that mimic the original (observed) data by preserving relationships between variables (Nowok et al., 2016) - may be useful in several areas such as healthcare, finance, data science, and machine learning (Dahmen & Cook, 2019; Kamthe et al., 2021; Nowok et al., 2016; Patki et al., 2016). As such, copula-based data generation models - probabilistic models that allow for the statistical properties of observed data to be modelled in terms of individual behavior and (inter-)dependencies (Joe, 2014) - have shown potential in several applications such as finance, data science, and meteorology (Kamthe et al., 2021; Li et al., 2020; Meyer, Nagler, et al., 2021; Patki et al., 2016). Although copula-based data generation tools have been developed for tabular data - e.g. the Synthetic Data Vault project using Gaussian copulas and generative adversarial networks (Patki et al., 2016; Xu & Veeramachaneni, 2018), or the Synthetic Data Generation via Gaussian Copula (Li et al., 2020) - in computational sciences such as weather and climate, data often consist of large, labelled multidimensional datasets with complex dependencies.

Here we introduce Synthia, an open-source multidimensional synthetic data generator written in Python for xarray’s (Hoyer & Hamman, 2017) labelled arrays and datasets with support for parametric and vine copulas models and functional principal component analysis (fPCA) - an extension of principal component analysis where data consist of functions instead of vectors (Ramsay & Silverman, 2005) - to allow for a wide range of data and dependent structures to be modelled. For efficiency, algorithms are implemented in NumPy (Harris et al., 2020) and SciPy (SciPy 1.0 Contributors et al., 2020) for Gaussian (parametric) copula and fPCA classes and rely on the C++ library vinecopulib (Nagler & Vatter, 2020b) through pyvinecopulib’s (Nagler & Vatter, 2020a) bindings for fast computation of vines.

Recent applications of Synthia include the generation of dependent (Meyer, Nagler, et al., 2021) and independent (Meyer, Hogan, et al., 2021) samples for improving the prediction of machine learning emulators in weather and climate. In this release we include examples and tutorials for univariate and multivariate synthetic data generation using copula and fPCA methods and look forward to enabling the generation of synthetic data in various scientific communities and for several applications.

Michaud, R. O. and Michaud, R. (2015). “Estimation Error and Portfolio Optimization: A Resampling Solution.” In: *SSRN e-Print*.

Markowitz (1959) mean-variance (MV) portfolio optimization has been the practical standard for asset allocation and equity portfolio management for almost fifty years. However, it is known to be overly sensitive to estimation error in risk-return estimates and have poor out-of-sample performance characteristics. The Resampled Efficiency (RE) techniques presented in Michaud (1998) introduce Monte Carlo methods to properly represent investment information uncertainty in computing MV portfolio optimality and in defining trading and monitoring rules. This paper reviews and updates the literature on estimation error and RE portfolio optimization and rebalancing. We resolve several open issues and misunderstandings that have emerged since Michaud (1998). In particular, we show RE optimization to be a Bayesian-based generalization and enhancement of Markowitz’s solution.

Mikalsen, K. O., Bianchi, F. M., Soguero-Ruiz, C., and Jenssen, R. (2018). “Time Series Cluster Kernel for Learning Similarities between Multivariate Time Series with Missing Data.” In: *Pattern Recognition* 76, pp. 569–581.

Similarity-based approaches represent a promising direction for time series analysis. However, many such methods rely on parameter tuning, and some have shortcomings if the time series are multivariate (MTS), due to dependencies between attributes, or the time series contain missing data. In this paper, we address these challenges within the powerful context of kernel methods by proposing the robust time series cluster kernel (TCK). The approach taken leverages the missing data handling properties of Gaussian mixture models (GMM) augmented with informative prior distributions. An ensemble learning approach is exploited to ensure robustness to parameters by combining the clustering results of many GMM to form the final kernel. We evaluate the TCK on synthetic and real data and compare to other state-of-the-art techniques. The experimental results demonstrate that the TCK is robust to parameter choices, provides competitive results for MTS without missing data and outstanding results for missing data.

Moudiki, T. and Planchet, F. (2016). “Economic Scenario Generators.” In: *Modelling in Life Insurance - A Management Perspective*. Ed. by J.-P. Laurent, R. Norberg, and F. Planchet. EAA Series. Springer International Publishing, pp. 81–104.

The projection of economic and financial risk factors is a key element of prospective analyzes made by life insurers, both for the calculation of reserves under Solvency 2 and for the asset allocation and management of financial risks. This projection is achieved in practice through - economic scenario generators- (ESG), which are inputs for the calculus of the economic value of assets and liabilities and the analysis of the distribution of this value. The calculation of economic values is based on the - no free lunch- assumption and therefore leads to model the risk factors in a riskneutral probability, while the analysis of the distribution of these values requires the projection of these factors under the historical probability. Therefore, the insurer must handle different representations of the risk factors, which requires looking at the characteristics of a risk neutral ESG, those of an - historical- one and the possible need for coherence between these two representations. This is what we propose to do in this chapter.

Nasri, B. R. and Rémillard, B. N. (2019). “Copula-based dynamic models for multivariate time series.” In: *Journal of Multivariate Analysis* 172, pp. 107–121.

In this paper, we propose an intuitive way to couple several dynamic time series models even when there are no innovations. This extends previous work for modeling dependence between innovations of stochastic volatility models. We consider time-dependent and time-independent copula models and we study the asymptotic behavior of some empirical processes constructed from pseudo-observations, as well as the behavior of maximum pseudo-likelihood estimators of the associated copula parameters. The results show that even if the univariate dynamic models depend on unknown parameters, the limiting behavior of many processes of interest does not depend on the estimation errors. One can perform tests for change points on the full distribution, the margins or the copula, as if the parameters of the dynamic models were known. This is also true for some parametric models of time-dependent copulas. This interesting property makes it possible to construct consistent tests of specification for the dependence models, without having to consider the dynamic time series models. Monte Carlo simulations are used to demonstrate the power of the proposed goodness-of-fit test in finite samples. An application to financial data is given.

Nasri, B. R., Rémillard, B. N., and Thioub, M. Y. (2020). “Goodness-of-fit for regime-switching copula models with application to option pricing.” In: *Canadian Journal of Statistics* 48(1), pp. 79–96.

We consider several time series, and for each of them, we fit an appropriate dynamic parametric model. This produces serially independent error terms for each time series. The dependence between these error terms is then modelled by a regime-switching copula. The EM algorithm is used for estimating the parameters and a sequential goodness-of-fit procedure based on Cramer-von Mises statistics is proposed to select the appropriate number of regimes. Numerical experiments are performed to assess the validity of the proposed methodology. As an example of application, we evaluate a European put-on-max option on the returns of two assets. To facilitate the use of our methodology, we have built a R package HMMcopula available on CRAN.

Naveau, P., Hannart, A., and Ribes, A. (2020). “Statistical methods for extreme event attribution in climate science.” In: *Annual Review of Statistics and its Application* 7(1), pp. 89–110.

Changes in the Earth’s climate have been increasingly observed. Assessing the likelihood that each of these changes has been caused by human influence is important for decision making on mitigation and adaptation policy. Because of their large societal and economic impacts, extreme events have garnered much media attentionave they become more frequent and more intense, and if so, why? To answer such questions, extreme event

attribution (EEA) tries to estimate extreme event likelihoods under different scenarios. Over the past decade, statistical methods and experimental designs based on numerical models have been developed, tested, and applied. In this article, we review the basic probability schemes, inference techniques, and statistical hypotheses used in EEA. To implement EEA analysis, the climate community relies on the use of large ensembles of climate model runs. We discuss, from a statistical perspective, how extreme value theory could help to deal with the different modeling uncertainties. In terms of interpretation, we stress that causal counterfactual theory offers an elegant framework that clarifies the design of event attributions. Finally, we pinpoint some remaining statistical challenges, including the choice of the appropriate spatio-temporal scales to enhance attribution power, the modeling of concomitant extreme events in a multivariate context, and the coupling of multi-ensemble and observational uncertainties.

Nguyen, H. D. and Chamroukhi, F. (2018). “[Practical and theoretical aspects of mixture-of-experts modeling: An overview.](#)” In: *WIREs Data Mining Knowl Discov* 8(4), e1246.

Mixture-of-experts (MoE) models are a powerful paradigm for modeling data arising from complex data generating processes (DGPs). In this article, we demonstrate how different MoE models can be constructed to approximate the underlying DGPs of arbitrary types of data. Due to the probabilistic nature of MoE models, we propose the maximum quasi-likelihood (MQL) approach as a method for estimating MoE model parameters from data, and we provide conditions under which MQL estimators are consistent and asymptotically normal. The blockwise minorization-maximization (blockwise-MM) algorithm framework is proposed as an all-purpose method for constructing algorithms for obtaining MQL estimators. An example derivation of a blockwise-MM algorithm is provided. We then present a method for constructing information criteria for estimating the number of components in MoE models and provide justification for the classic Bayesian information criterion (BIC). We explain how MoE models can be used to conduct classification, clustering, and regression and illustrate these applications via two worked examples.

Ni, H., Szpruch, L., Wiese, M., Liao, S., and Xiao, B. (2020). “[Conditional Sig-Wasserstein GANs for Time Series Generation.](#)” In: *arXiv e-Print*.

Generative adversarial networks (GANs) have been extremely successful in generating samples, from seemingly high dimensional probability measures. However, these methods struggle to capture the temporal dependence of joint probability distributions induced by time-series data. Furthermore, long time-series data streams hugely increase the dimension of the target space, which may render generative modeling infeasible. To overcome these challenges, we integrate GANs with mathematically principled and efficient path feature extraction called the signature of a path. The signature of a path is a graded sequence of statistics that provides a universal description for a stream of data, and its expected value characterizes the law of the time-series model. In particular, we develop a new metric, (conditional) Sig- $W_1$ , that captures the (conditional) joint law of time series models, and use it as a discriminator. The signature feature space enables the explicit representation of the proposed discriminators which alleviates the need for expensive training. Furthermore, we develop a novel generator, called the conditional AR-FNN, which is designed to capture the temporal dependence of time series and can be efficiently trained. We validate our method on both synthetic and empirical datasets and observe that our method consistently and significantly outperforms state-of-the-art benchmarks with respect to measures of similarity and predictive ability.

Niemann, J.-H., Klus, S., and Schutte, C. (2021). “[Data-driven model reduction of agent-based systems using the Koopman generator.](#)” In: *arXiv e-Print*.

The dynamical behavior of social systems can be described by agent-based models. Although single agents follow easily explainable rules, complex time-evolving patterns emerge due to their interaction. The simulation and analysis of such agent-based models, however, is often prohibitively time-consuming if the number of agents is large. In this paper, we show how Koopman operator theory can be used to derive reduced models of agent-based systems using only simulation data. Our goal is to learn coarse-grained models and to represent the reduced dynamics by ordinary or stochastic differential equations. The new variables are, for instance, aggregated state variables of the agent-based model, modeling the collective behavior of larger groups or the entire population. Using benchmark problems with known coarse-grained models, we demonstrate that the obtained reduced systems are in good agreement with the analytical results, provided that the numbers of agents is sufficiently large.

Oneto, L. (2019). “[Resampling Methods.](#)” In: *Model Selection and Error Estimation in a Nutshell*. Springer International Publishing, pp. 25–31.



Resampling methods, also called Out-of-Sample methods, are favoured by practitioners because they work well in many situations and allow the application of simple statistical techniques for estimating the quantities of interest.

Opdyke, J. D. (2020). “Full Probabilistic Control for Direct and Robust, Generalized and Targeted Stressing of the Correlation Matrix (Even When Eigenvalues are Empirically Challenging.” In: *QuantMinds/RiskMinds Conference*.

In practice, risk measurement of the majority of enterprise-level portfolios, and even many investment portfolios, requires stressing the correlation matrix directly, rather than (solely) stressing its underlying variables, due to 1. data paucity or incomplete time series, 2. matrices where at least some of the values are based on subject-matter expertise, and/or 3. the need to specify and test the effects of correlation values nowhere close to those reflected in historical data (i.e. the majority of extreme scenarios used in forward-looking stress testing). Surprisingly few papers in the literature address this common, real-world situation, and their approaches arguably are either ad-hoc, lack solid statistical underpinnings, do not allow for direct, probability-based stressing, and/or remain non-robust as they fail under empirically challenging conditions (e.g. near-zero eigenvalues resulting from the need to first enforce positive definiteness of the matrix). We borrow from recent advances in the literature for generating random correlation matrices (based on the identity matrix) to design a method that both mitigates and eliminates these drawbacks when directly stressing real-world correlation matrices (other than the identity matrix). Our approach can be used for both generalized and targeted stressing. The former perturbs the entire correlation matrix, which can be used to account for difficult-to-model or difficult-to-anticipate second and third order effects of extreme scenarios, as well as providing much needed percentiles of the distribution of the entire matrix. Targeted stressing, on the other hand, allows for particular correlation values to be changed by specified amounts based directly on the probability of observing such changes due to the event/scenario. And both generalized and targeted stressing can be performed concurrently, based on the same proposed approach, which provides full probabilistic control while automatically enforcing positive definiteness. We demonstrate the method on realistic, reasonably large matrices (100x100) that have had positive definiteness enforced via Higham(2002), reflecting a common occurrence for most enterprise-level portfolios and even many investment portfolios. Although it requires numeric integration for all but very small matrices, our approach’s runtimes are comparable to those of competing methods. Implementation is straightforward, and results robustly outperform existing methods in the literature, especially when matrices are empirically challenging (e.g. near-zero eigenvalues).

Ortec Finance (2014). “Ex ante risk management with scenarios.” In: *CFA Society Switzerland*.

Ex-ante risk management with help of scenarios including important information such as the impact of the business cycle is an innovative methodology used by sophisticated managers gaining worldwide popularity. Advanced stochastic scenario analysis supports financial decision-making when dealing with uncertainty through accounting for structural changes impacting trend values as well as short term events influencing the market momentum. The extent to which investor objectives will be achieved without violating risk limits is influenced by the path that financial and economic risk drivers will follow. A frequency domain based approach differs from traditional simulation approaches through the creation of plausible and relevant scenarios. The Dynamic Scenario Generator (DSG) employs a unique combination of statistical and econometric techniques. The methodology is a mixture of frequency domain techniques, dynamic factor models and special approaches for dealing with non-normal distribution characteristics such as skewness and fat tails. The main purpose of the central frequency domain methodology is to describe all empirical features ( stylized facts ) of the time series behavior of financial and economic variables simultaneously, rather than focusing on only one or a few aspects in isolation.

Packham, N. and Woebbecking, F. (2021). “Correlation scenarios and correlation stress testing.” In: *SSRN e-Print*.

We develop a general approach for stress testing correlations of financial asset portfolios. The correlation matrix of asset returns is specified in a parametric form, where correlations are represented as a function of risk factors, such as country and industry factors. A sparse factor structure linking assets and risk factors is built using Bayesian variable selection methods. Regular calibration yields a joint distribution of economically meaningful stress scenarios of the factors. As such, the method also lends itself as a reverse stress testing framework: using the Mahalanobis distance or highest density regions (HDR) on the joint risk factor distribution allows to infer worst-case correlation scenarios. We give examples of stress tests on a large portfolio of European and North American stocks.

Padhi, I., Schiff, Y., Melnyk, I., Rigotti, M., Mroueh, Y., Dognin, P., Ross, J., Nair, R., and Altman, E. (2021). “Tabular Transformers for Modeling Multivariate Time Series.” In: *arXiv e-Print*.

Tabular datasets are ubiquitous in data science applications. Given their importance, it seems natural to apply state-of-the-art deep learning algorithms in order to fully unlock their potential. Here we propose neural network models that represent tabular time series that can optionally leverage their hierarchical structure. This results in two architectures for tabular time series: one for learning representations that is analogous to BERT and can be pre-trained end-to-end and used in downstream tasks, and one that is akin to GPT and can be used for generation of realistic synthetic tabular sequences. We demonstrate our models on two datasets: a synthetic credit card transaction dataset, where the learned representations are used for fraud detection and synthetic data generation, and on a real pollution dataset, where the learned encodings are used to predict atmospheric pollutant concentrations. Code and data are available at <https://github.com/IBM/TabFormer>.

Page, S. (2013). “How to combine long and short return histories efficiently.” In: *Financial Analysts Journal* 69(1), pp. 45–52.

A common challenge in portfolio risk analysis is that certain assets have shorter return histories than others. Unfortunately, many standard portfolio risk analysis techniques—including historical tail risk measurement, regime-dependent risk analysis, and bootstrapping simulations—require full return histories for all assets or risk factors. The author presents easy instructions on how to efficiently combine data for investments whose histories differ in length and offers a new model to better account for non-normal distributions.

Pan, I., Mason, L., and Matar, O. (2021). “Data-Centric Engineering: integrating simulation, machine learning and statistics. Challenges and Opportunities.” In: *arXiv e-Print*.

Recent advances in machine learning, coupled with low-cost computation, availability of cheap streaming sensors, data storage and cloud technologies, has led to widespread multi-disciplinary research activity with significant interest and investment from commercial stakeholders. Mechanistic models, based on physical equations, and purely data-driven statistical approaches represent two ends of the modelling spectrum. New hybrid, data-centric engineering approaches, leveraging the best of both worlds and integrating both simulations and data, are emerging as a powerful tool with a transformative impact on the physical disciplines. We review the key research trends and application scenarios in the emerging field of integrating simulations, machine learning, and statistics. We highlight the opportunities that such an integrated vision can unlock and outline the key challenges holding back its realisation. We also discuss the bottlenecks in the translational aspects of the field and the long-term upskilling requirements of the existing workforce and future university graduates.

Pancheekha, A., Tull, R., and Bell, M. M. (2018). “Ensemble Active Management.” In: *SSRN e-Print*.

This White Paper questions the superiority of the traditional Active Management paradigm. Do stand-alone, -expert investment managers or management teams, with well-defined yet rigidly entrenched philosophies and methodologies, deliver optimal results? The conclusion, derived from a database reflecting 30,000 test portfolios and 165 million data points, was that they do not.

Papenbrock, J., Schwendner, P., Jaeger, M., and Krugel, S. (2021). “Matrix Evolutions: Synthetic Correlations and Explainable Machine Learning for Constructing Robust Investment Portfolios.” In: *The Journal of Financial Data Science* 3(2), pp. 51–69.

In this article, the authors present a novel and highly flexible concept to simulate correlation matrixes of financial markets. It produces realistic outcomes regarding stylized facts of empirical correlation matrixes and requires no asset return input data. The matrix generation is based on a multiobjective evolutionary algorithm, so the authors call the approach matrix evolutions. It is suitable for parallel implementation and can be accelerated by graphics processing units and quantum-inspired algorithms. The approach is useful for backtesting, pricing, and hedging correlation-dependent investment strategies and financial products. Its potential is demonstrated in a machine learning case study for robust portfolio construction in a multi-asset universe: An explainable machine learning program links the synthetic matrixes to the portfolio volatility spread of hierarchical risk parity versus equal risk contribution.

Papoudakis, G., Christianos, F., and Albrecht, S. V. (2021). “Agent Modelling under Partial Observability for Deep Reinforcement Learning.” In: *arXiv e-Print*.

Modelling the behaviours of other agents is essential for understanding how agents interact and making effective decisions. Existing methods for agent modelling commonly assume knowledge of the local observations and chosen actions of the modelled agents during execution. To eliminate this assumption, we extract representations from the local information of the controlled agent using encoder-decoder architectures. Using the observations and actions of the modelled agents during training, our models learn to extract representations about the modelled agents conditioned only on the local observations of the controlled agent. The representations are used to augment the controlled agent’s decision policy which is trained via deep reinforcement learning; thus, during execution,

the policy does not require access to other agents’ information. We provide a comprehensive evaluation and ablations studies in cooperative, competitive and mixed multi-agent environments, showing that our method achieves higher returns than baseline methods which do not use the learned representations.

Pathak, P. K. and Rao, C. R. (2013). “[The Sequential Bootstrap](#).” In: *Handbook of Statistics*. Vol. 31. Elsevier, pp. 2–18.

The main focus of this article is to carefully examine the information content of resampling techniques in bootstrap from a survey sampling point of view. Given an observed sample of size  $n$ , resampling for bootstrap involves repeated trials of simple random sampling with replacement (SRSWR). It is well known that SRSWR does not result in samples that are equally informative (Pathak, 1964). In 1997, Rao et al. introduced a sequential bootstrap resampling scheme in the literature, stemming from the observation made by Efron (1983) that the usual bootstrap samples are supported on average on approximately 0.632 of the original data points. The primary goal of our sequential bootstrap was to stabilize the information content of bootstrapped samples as well as to investigate whether consistency and second-order correctness of Efron’s simple bootstrap carried over to this sequential case. In Rao et al. (1997), we showed that the main empirical characteristics of the sequential bootstrap asymptotically differ from those of the corresponding simple bootstrap by a magnitude of  $\sqrt{n}$ . In Babu et al. (1999), we established the second-order correctness of the sequential bootstrap for a general class of estimators under regularity conditions analogous to those for the simple bootstrap. In a separate paper, Shoemaker and Pathak (2001) carried out simulations similar to those of Mammen (1992), confirming the consistency of the sequential bootstrap. More recently, Jimenez-Gamero et al. (2006), Pino-Mejias et al. (2010), Pauly (2011), and others have investigated the sequential bootstrap and its variants both empirically as well as theoretically. These investigations indicate that the sequential bootstrap is likely to perform better than the simple bootstrap in small to moderate samples. We present an expository account of these findings and discuss their potential applications.

Paul, J., Michael, B.-S., Pedro, M., Rajbir, S. N., Shubham, K., Valentin, F., Jan, G., and Tim, J. (2022). “[PSA-GAN: Progressive Self Attention GANs for Synthetic Time Series](#).” In: *arXiv e-Print*.

Realistic synthetic time series data of sufficient length enables practical applications in time series modeling tasks, such as forecasting, but remains a challenge. In this paper we present PSA-GAN, a generative adversarial network (GAN) that generates long time series samples of high quality using progressive growing of GANs and self-attention. We show that PSA-GAN can be used to reduce the error in two downstream forecasting tasks over baselines that only use real data. We also introduce a Frechet-Inception Distance-like score, Context-FID, assessing the quality of synthetic time series samples. In our downstream tasks, we find that the lowest scoring models correspond to the best-performing ones. Therefore, Context-FID could be a useful tool to develop time series GAN models.

Pei, H., Ren, K., Yang, Y., Liu, C., Qin, T., and Li, D. (2021). “[Towards Generating Real-World Time Series Data](#).” In: *arXiv e-Print*.

Time series data generation has drawn increasing attention in recent years. Several generative adversarial network (GAN) based methods have been proposed to tackle the problem usually with the assumption that the targeted time series data are well-formatted and complete. However, real-world time series (RTS) data are far away from this utopia, e.g., long sequences with variable lengths and informative missing data raise intractable challenges for designing powerful generation algorithms. In this paper, we propose a novel generative framework for RTS data - RTSGAN to tackle the aforementioned challenges. RTSGAN first learns an encoder-decoder module which provides a mapping between a time series instance and a fixed-dimension latent vector and then learns a generation module to generate vectors in the same latent space. By combining the generator and the decoder, RTSGAN is able to generate RTS which respect the original feature distributions and the temporal dynamics. To generate time series with missing values, we further equip RTSGAN with an observation embedding layer and a decide-and-generate decoder to better utilize the informative missing patterns. Experiments on the four RTS datasets show that the proposed framework outperforms the previous generation methods in terms of synthetic data utility for downstream classification and prediction tasks.

Perez, J., Arroba, P., and Moya, J. M. (2022). “[Data augmentation through multivariate scenario forecasting in Data Centers using Generative Adversarial Networks](#).” In: *arXiv e-Print*.

The Cloud paradigm is at a critical point in which the existing energy-efficiency techniques are reaching a plateau, while the computing resources demand at Data Center facilities continues to increase exponentially. The main challenge in achieving a global energy efficiency strategy based on Artificial Intelligence is that we need massive amounts of data to feed the algorithms. Nowadays, any optimization strategy must begin with

data. However, companies with access to these large amounts of data decide not to share them because it could compromise their security. This paper proposes a time-series data augmentation methodology based on synthetic scenario forecasting within the Data Center. For this purpose, we will implement a powerful generative algorithm: Generative Adversarial Networks (GANs). The use of GANs will allow us to handle multivariate data and data from different natures (e.g., categorical). On the other hand, adapting Data Centers' operational management to the occurrence of sporadic anomalies is complicated due to the reduced frequency of failures in the system. Therefore, we also propose a methodology to increase the generated data variability by introducing on-demand anomalies. We validated our approach using real data collected from an operating Data Center, successfully obtaining forecasts of random scenarios with several hours of prediction. Our research will help to optimize the energy consumed in Data Centers, although the proposed methodology can be employed in any similar time-series-like problem.

Perumal, R. and van Zyl, T. L. (2020). "Surrogate Assisted Methods for the Parameterisation of Agent-Based Models." In: *arXiv e-Print*.

Parameter calibration is a major challenge in agent-based modelling and simulation (ABMS). As the complexity of agent-based models (ABMs) increase, the number of parameters required to be calibrated grows. This leads to the ABMS equivalent of the curse of dimensionality. We propose an ABMS framework which facilitates the effective integration of different sampling methods and surrogate models (SMs) in order to evaluate how these strategies affect parameter calibration and exploration. We show that surrogate assisted methods perform better than the standard sampling methods. In addition, we show that the XGBoost and Decision Tree SMs are most optimal overall with regards to our analysis.

Pesenti, S. M., Bettini, A., Millossovich, P., and Tsanakas, A. (2020). "Scenario weights for importance measurement (SWIM) - an R package for sensitivity analysis." In: *SSRN e-Print*.

The SWIM package implements a flexible sensitivity analysis framework, based primarily on results and tools developed by Pesenti et al. (2019). SWIM provides a stressed version of a stochastic model, subject to model components (random variables) fulfilling given probabilistic constraints (stresses). Possible stresses can be applied on moments, probabilities of given events, and risk measures such as Value-at-Risk and Expected Shortfall. SWIM operates upon a single set of simulated scenarios from a stochastic model, returning scenario weights, which encode the required stress and allow monitoring the impact of the stress on all model components. The scenario weights are calculated to minimise the relative entropy with respect to the baseline model, subject to the stress applied. As well as calculating scenario weights, the package provides tools for the analysis of stressed models, including plotting facilities and evaluation of sensitivity measures. SWIM does not require additional evaluations of the simulation model or explicit knowledge of its underlying statistical and functional relations; hence it is suitable for the analysis of black box models. The capabilities of SWIM are demonstrated through a case study of a credit portfolio model.

Pesenti, S. M., Millossovich, P., and Tsanakas, A. (2019). "Reverse sensitivity testing: What does it take to break the model?" In: *European Journal of Operational Research* 274(2), pp. 654–670.

Sensitivity analysis is an important component of model building, interpretation and validation. A model comprises a vector of random input factors, an aggregation function mapping input factors to a random output, and a (baseline) probability measure. A risk measure, such as Value-at-Risk and Expected Shortfall, maps the distribution of the output to the real line. As is common in risk management, the value of the risk measure applied to the output is a decision variable. Therefore, it is of interest to associate a critical increase in the risk measure to specific input factors. We propose a global and model-independent framework, termed sensitivity testing, comprising three steps: (a) an output stress is specified, corresponding to an increase in the risk measure(s); (b) a (stressed) probability measure is derived, minimising the Kullback-Leibler divergence with respect to the baseline probability, under constraints generated by the output stress; (c) changes in the distributions of input factors are evaluated. We argue that a substantial change in the distribution of an input factor corresponds to high sensitivity to that input and introduce a novel sensitivity measure to formalise this insight. Implementation of reverse sensitivity testing in a Monte-Carlo setting can be performed on a single set of input/output scenarios, simulated under the baseline model. Thus the approach circumvents the need for additional computationally expensive evaluations of the aggregation function. We illustrate the proposed approach through a numerical example of a simple insurance portfolio and a model of a London Insurance Market portfolio used in industry.

Qi, D. and Majda, A. J. (2019). "Using machine learning to predict extreme events in complex systems." In: *Proceedings of the National Academy of Sciences of the United States of America*.

Extreme events and the related anomalous statistics are ubiquitously observed in many natural systems, and the development of efficient methods to understand and accurately predict such representative features remains a grand challenge. Here, we investigate the skill of deep learning strategies in the prediction of extreme events in complex turbulent dynamical systems. Deep neural networks have been successfully applied to many imaging processing problems involving big data, and have recently shown potential for the study of dynamical systems. We propose to use a densely connected mixed-scale network model to capture the extreme events appearing in a truncated Korteweg-de Vries (tKdV) statistical framework, which creates anomalous skewed distributions consistent with recent laboratory experiments for shallow water waves across an abrupt depth change, where a remarkable statistical phase transition is generated by varying the inverse temperature parameter in the corresponding Gibbs invariant measures. The neural network is trained using data without knowing the explicit model dynamics, and the training data are only drawn from the near-Gaussian regime of the tKdV model solutions without the occurrence of large extreme values. A relative entropy loss function, together with empirical partition functions, is proposed for measuring the accuracy of the network output where the dominant structures in the turbulent field are emphasized. The optimized network is shown to gain uniformly high skill in accurately predicting the solutions in a wide variety of statistical regimes, including highly skewed extreme events. The technique is promising to be further applied to other complicated high-dimensional systems.

Raab, G. M., Nowok, B., and Dibben, C. (2021). “Assessing, visualizing and improving the utility of synthetic data.” In: *arXiv e-Print*.

A number of measures have been proposed to assess the utility of the synthetic data. These include measures based on distances between the two distributions and others based on combining the original and synthetic data and predicting the origin with a propensity score. The methods will be reviewed and compared, and relations between them illustrated. These measures are incorporated into utility modules in the synthpop package that include methods to visualize the results. We illustrate how to compare different syntheses and to diagnose which aspect of the synthetic data differs from the original. The utility functions were originally designed to be used for synthetic data objects of class `synds`, created by synthpop, but they can now be used to compare synthetic data created by other methods with the original records. The utility measures can be standardized by their expected Null distributions from a correct synthesis model. If they are used to evaluate other types of altered data, not generated from a model, then this standardisation can be interpreted as giving the ratio of the difference for the original to the expected stochastic error.

Racca, A. and Magri, L. (2022). “Statistical prediction of extreme events from small datasets.” In: *arXiv e-Print*.

We propose Echo State Networks (ESNs) to predict the statistics of extreme events in a turbulent flow. We train the ESNs on small datasets that lack information about the extreme events. We assess whether the networks are able to extrapolate from the small imperfect datasets and predict the heavy-tail statistics that describe the events. We find that the networks correctly predict the events and improve the statistics of the system with respect to the training data in almost all cases analysed. This opens up new possibilities for the statistical prediction of extreme events in turbulence.

Raghunathan, T. E. (2021). “Synthetic Data.” In: *Annual Review of Statistics and Its Application* 8(1), pp. 129–140.

Demand for access to data, especially data collected using public funds, is ever growing. At the same time, concerns about the disclosure of the identities of and sensitive information about the respondents providing the data are making the data collectors limit the access to data. Synthetic data sets, generated to emulate certain key information found in the actual data and provide the ability to draw valid statistical inferences, are an attractive framework to afford widespread access to data for analysis while mitigating privacy and confidentiality concerns. The goal of this article is to provide a review of various approaches for generating and analyzing synthetic data sets, inferential justification, limitations of the approaches, and directions for future research.

Raileanu, R., Goldstein, M., Yarats, D., Kostrikov, I., and Fergus, R. (2021). “Automatic Data Augmentation for Generalization in Deep Reinforcement Learning.” In: *arXiv e-Print*.

Deep reinforcement learning (RL) agents often fail to generalize to unseen scenarios, even when they are trained on many instances of semantically similar environments. Data augmentation has recently been shown to improve the sample efficiency and generalization of RL agents. However, different tasks tend to benefit from different kinds of data augmentation. In this paper, we compare three approaches for automatically finding an appropriate augmentation. These are combined with two novel regularization terms for the policy and value function, required to make the use of data augmentation theoretically sound for certain actor-critic algorithms. We evaluate our methods on the Procgen benchmark which consists of 16 procedurally-generated environments and show that it improves test performance by 40% relative to standard RL algorithms. Our agent outperforms other baselines



specifically designed to improve generalization in RL. In addition, we show that our agent learns policies and representations that are more robust to changes in the environment that do not affect the agent, such as the background. Our implementation is available at <https://github.com/rraileanu/auto-drac>.

Rajabi, A. and Garibay, O. O. (2021). “TabFairGAN: Fair Tabular Data Generation with Generative Adversarial Networks.” In: *arXiv e-Print*.

With the increasing reliance on automated decision making, the issue of algorithmic fairness has gained increasing importance. In this paper, we propose a Generative Adversarial Network for tabular data generation. The model includes two phases of training. In the first phase, the model is trained to accurately generate synthetic data similar to the reference dataset. In the second phase we modify the value function to add fairness constraint, and continue training the network to generate data that is both accurate and fair. We test our results in both cases of unconstrained, and constrained fair data generation. In the unconstrained case, i.e. when the model is only trained in the first phase and is only meant to generate accurate data following the same joint probability distribution of the real data, the results show that the model beats state-of-the-art GANs proposed in the literature to produce synthetic tabular data. Also, in the constrained case in which the first phase of training is followed by the second phase, we train the network and test it on four datasets studied in the fairness literature and compare our results with another state-of-the-art pre-processing method, and present the promising results that it achieves. Comparing to other studies utilizing GANs for fair data generation, our model is comparably more stable by using only one critic, and also by avoiding major problems of original GAN model, such as mode-dropping and non-convergence, by implementing a Wasserstein GAN.

Rebonato, R. (2018). “The quickest way to lose the money you cannot afford to lose: reverse stress testing with maximum entropy.” In: *Journal of Risk* 20(3), pp. 83–93.

We extend a technique devised by Saroka and Rebonato to “optimally” deform a yield curve in order to deal with a common and practically relevant class of optimization problems subject to linear constraints. In particular, we show how the idea can be applied to the case of reverse stress testing, and we present a case study to illustrate how it works. Finally, we point out a maximum-entropy interpretation of (or justification for) the procedure and present some obvious generalizations.

Rikli, S., Bigler, D. N., Pfenninger, M., and Osterrieder, J. (2021). “Wasserstein GAN: Deep Generation applied on Bitcoins financial time series.” In: *arXiv e-Print*.

Modeling financial time series is challenging due to their high volatility and unexpected happenings on the market. Most financial models and algorithms trying to fill the lack of historical financial time series struggle to perform and are highly vulnerable to overfitting. As an alternative, we introduce in this paper a deep neural network called the WGAN-GP, a data-driven model that focuses on sample generation. The WGAN-GP consists of a generator and discriminator function which utilize an LSTM architecture. The WGAN-GP is supposed to learn the underlying structure of the input data, which in our case, is the Bitcoin. Bitcoin is unique in its behavior; the prices fluctuate what makes guessing the price trend hardly impossible. Through adversarial training, the WGAN-GP should learn the underlying structure of the bitcoin and generate very similar samples of the bitcoin distribution. The generated synthetic time series are visually indistinguishable from the real data. But the numerical results show that the generated data were close to the real data distribution but distinguishable. The model mainly shows a stable learning behavior. However, the model has space for optimization, which could be achieved by adjusting the hyperparameters.

Rocca, J. and Rocca, B. (2019). “Intuitively Understanding Variational Autoencoders.” In: *Towards Data Science*.

Rojas, H. and Dias, D. (2021). “Stress testing network reconstruction via graphical causal model.” In: *Applied Stochastic Models in Business and Industry*.

An resilience optimal evaluation of financial portfolios implies having plausible hypotheses about the multiple interconnections between the macroeconomic variables and the risk parameters. In this article, we propose a graphical model for the reconstruction of the causal structure that links the multiple macroeconomic variables and the assessed risk parameters, it is this structure that we call stress testing network. In this model, the relationships between the macroeconomic variables and the risk parameter define a “relational graph” among their time-series, where related time-series are connected by an edge. Our proposal is based on the temporal causal models, but unlike, we incorporate specific conditions in the structure which correspond to intrinsic characteristics this type of networks. Using the proposed model and given the high-dimensional nature of the problem, we used regularization methods to efficiently detect causality in the time-series and reconstruct the underlying causal structure. In addition, we illustrate the use of model in credit risk data of a portfolio. Finally, we discuss its uses and practical benefits in stress testing.

Romano, J. P. and Wolf, M. (2018). “Multiple Testing of One-Sided Hypotheses: Combining Bonferroni and the Bootstrap.” In: *Predictive econometrics and big data*. Ed. by V. Kreinovich, S. Sriboonchitta, and N. Chakpitak. Vol. 753. Springer International Publishing, pp. 78–94.

In many multiple testing problems, the individual null hypotheses (i) concern univariate parameters and (ii) are one-sided. In such problems, power gains can be obtained for bootstrap multiple testing procedures in scenarios where some of the parameters are ‘deep in the null’ by making certain adjustments to the null distribution under which to resample. In this paper, we compare a Bonferroni adjustment that is based on finite-sample considerations with certain ‘asymptotic’ adjustments previously suggested in the literature.

Rosen, D. (2015a). “Integrating Economic Scenarios with Advanced Scenario Analytics to Manage Investment Portfolios.” In: *SSRN e-Print*.

Scenarios are the language of Risk. The quality of a risk management analysis depends on the ability to generate relevant forward-looking scenarios that properly represent the future. While the future cannot be predicted, the combination of economic and expert analysis, with advanced scenario and portfolio analytics, provides a strong basis for managing risk and making better, more informed investment decisions. In this workshop, we present the latest simulation methodologies to analyze in practice the risk and investment opportunities of financial portfolios. We show how economic forecasts and expert views can be effectively combined with advanced Monte Carlo and Conditional Scenario methods to create meaningful scenarios for buy-side investment analysis, stress testing of credit portfolios, and systemic risk. Finally, we illustrate their practical application through various real-life examples where we analyze the performance of portfolios under recent economic research reports, regulatory scenarios and expert subjective views.

Rosen, D. (2015b). “Re-Thinking Scenarios: Stress Testing of Multi-Asset Portfolios by Integrating Economic Scenarios with Advanced Simulation Analytics.” In: *SSRN e-Print*.

Scenarios are the language of Risk. While scenario analysis and stress testing have been an explicit part of risk management methodologies and systems for over two decades, the typical scenario and stress testing tools haven’t evolved much and are still generally quite static and largely subjective. In this talk, we present a simple and powerful approach to create meaningful stress scenarios for risk management and investment analysis of multi-asset portfolios, which effectively combines economic forecasts and expert views with portfolio simulation methods. Expert scenarios are typically described in terms of a small number of key economic variables or factors. However, when applied to a portfolio, they are incomplete: they generally do not describe what occurs to all relevant market risk factors that affect the portfolio. We need to understand how these market risk factors behave, conditional on the outcome of the economic factors. The key insight to our approach is that the conditional expectation, and more generally the full conditional distribution of all the factors, and of the portfolio P and L, can be estimated directly from a pre-computed simulation using Least Squares Regression. We refer to this approach as Least Squares Stress Testing (LSST). LSST is a simulation-based conditional scenario generation method that offers many advantages over more traditional analytical methods. Simulation techniques are simple, flexible, and provide very transparent results, which are auditable and easy to explain. LSST can be applied to both market and credit risk stress testing with a large number of risk factors, which can follow completely general stochastic processes, with fat-tails, non-parametric and general codependence structures, autocorrelation, etc. LSST further produces explicit risk factor P and L contributions. From a methodology perspective, we also discuss some of the assumptions the LSST approach, statistical tests to check when these assumptions fail, and remedies that can be applied. Finally, we illustrate the application of the methodology through the analysis of the performance of a real-life portfolio under scenarios from a recent economic research report as well as regulatory scenarios.

Rosen, D. and Saunders, D. (2015). “Regress Under Stress: A Simple Least-Squares Method for Integrating Economic Scenarios with Risk Simulations.” In: *SSRN e-Print*.

Scenarios are the language of Risk. While scenario analysis and stress testing have been an explicit part of risk management methodologies and systems for over two decades, the typical scenario and stress testing tools haven’t evolved much and are still generally quite static and largely subjective. In this paper, we present a simple and powerful approach to create meaningful stress scenarios for risk management and investment analysis of multi-asset portfolios, which effectively combines economic forecasts and expert views with portfolio simulation methods. Expert scenarios are typically described in terms of a small number of key economic variables or factors. However, when applied to a portfolio, they are incomplete they generally do not describe what occurs to all relevant market risk factors that affect the portfolio. We need to understand how these market risk factors behave, conditional on the outcome of the economic factors. The key insight to our approach is that the

conditional expectation, or more generally the full conditional distribution of all the factors, and of the portfolio P and L, can be estimated directly from a pre-computed simulation using Least Squares Regressions (LSR). All the conditional scenario analytics can be derived from the regression results. We refer to this approach as Least Squares Stress Testing (LSST). LSST is a simulation-based conditional scenario generation method that offers many advantages over more traditional analytical methods. Simulation techniques are simple, flexible, and provide very transparent results, which are auditable and easy to explain. LSST can be applied to a large number of risk factors, which can follow completely general joint stochastic processes, with fat-tails, non-parametric and general codependence structures, autocorrelation, etc. LSST further produces explicit risk factor P and L contributions. We illustrate the application of the methodology through the analysis of the performance of a real-life portfolio under recent economic research reports.

Rosen, D. and Saunders, D. (2016). “Regress under stress: A simple least-squares method for integrating economic scenarios with risk simulations.” In: *Journal of Risk Management in Financial Institutions* 9(4), pp. 391–412.

We present a simple and powerful approach to create meaningful stress scenarios for risk management and investment analysis of multi-asset portfolios, which effectively combines economic forecasts and ‘expert’ views with portfolio simulation methods. Expert scenarios are typically described in terms of a small number of key economic variables or factors. However, when applied to a portfolio, they are incomplete - they generally do not describe what occurs to all relevant market risk factors that affect the portfolio. We need to understand how these market risk factors behave, conditional on the outcome of the economic factors. The key insight to our approach is that the conditional expectation, and more generally the full conditional distribution of all the factors, and of the portfolio profit and loss (P and L), can be estimated directly from a pre-computed simulation using least squares regression. We refer to this approach as least squares stress testing (LSST). LSST is a simulation-based conditional scenario generation method that offers many advantages over more traditional analytical methods. Simulation techniques are simple, flexible and provide very transparent results, which are auditable and easy to explain. LSST can be applied to both market and credit risk stress testing with a large number of risk factors, which can follow completely general stochastic processes, with fat-tails, non-parametric and general co-dependence structures, autocorrelation, etc. LSST further produces explicit risk factor P and L contributions. We demonstrate the methodology in detail with the practical example of a multi-asset investment portfolio and economic scenarios from an industry report.

Rosolia, A. and Osterrieder, J. (2021). “Analyzing Deep Generated Financial Time Series for Various Asset Classes.” In: *SSRN e-Print*.

Generative Adversarial Networks (GANs) have shown remarkable success as a framework for training models to produce realistic-looking data. In this work, we propose a GAN to produce realistic real-valued time series, with an emphasis on their application to financial data. Our aim is having a GAN, applied on various financial time series for various asset classes, that can reflect all the characteristics of them, as well as the characteristics we may be unaware of, as GANs learn the underlying structure of our data, rather than just a set of features. If we are able to achieve this, the synthetic datasets we create could be used for a variety of purposes including model training and model selection. In this paper we try to train a GAN with real data, representing one asset of different asset classes such as commodities, forex, futures, index and shares.

Ruenzi, S., Ungeheuer, M., and Weigert, F. (2020). “Joint Extreme events in equity returns and liquidity and their cross-sectional pricing implications.” In: *Journal of Banking & Finance* 115, p. 105809.

We merge the literature on downside return risk and liquidity risk and introduce the concept of extreme downside liquidity (EDL) risks. The cross-section of stock returns reflects a premium if a stock’s return (liquidity) is lowest at the same time when the market liquidity (return) is lowest. This effect is not driven by linear or downside liquidity risk or extreme downside return risk and is mainly driven by more recent years. There is no premium for stocks whose liquidity is lowest when market liquidity is lowest.

Sahamkhadam, M. and Stephan, A. (2021). “Portfolio Optimization Based on Forecasting Models Using Vine Copulas: An Empirical Assessment for the Financial Crisis.” In: *SSRN e-Print*.

We employ and examine vine copulas in modeling symmetric and asymmetric dependency structures and forecasting financial returns. We analyze the asset allocations performed during the 2008-2009 global financial crisis and the year 2020 COVID-19 crisis and test different portfolio strategies such as maximum Sharpe ratio, minimum variance, and minimum conditional Value-at-Risk. Using international financial markets, we specify the regular, drawable, and canonical vine copulas, such as the Student-t, Clayton, Frank, Joe, Gumbel, and mixed copulas, and analyze both in-sample and out-of-sample portfolio performances. Out-of-sample portfolio back-testing shows that vine copulas reduce portfolio risk better than simple copulas. Overall, we find that the

Student-t drawable vine copula models perform best with regard to risk reduction, both for the entire period 2005-2012 as well as during the global financial crisis.

Salazar, A., Vergara, L., and Safont, G. (2021). “Generative Adversarial Networks and Markov Random Fields for oversampling very small training sets.” In: *Expert Systems with Applications* 163, p. 113819.

In this work, we propose a new method for oversampling the training set of a classifier, in a scenario of extreme scarcity of training data. It is based on two concepts: Generative Adversarial Networks (GAN) and vector Markov Random Field (vMRF). Thus, the generative block of GAN uses the vMRF model to synthesize surrogates by the Graph Fourier Transform. Then, the discriminative block implements a linear discriminant on features measuring clique similarities between the synthesized and the original instances. Both blocks iterate until the linear discriminant cannot discriminate the synthetic from the original instances. We have assessed the new method, called Generative Adversarial Network Synthesis for Oversampling (GANSO), with both simulated and real data in experiments where the classifier is to be trained with just 3 or 5 instances. The applications consisted of classification of stages of neuropsychological tests using electroencephalographic (EEG) and functional magnetic resonance imaging (fMRI) data and classification of sleep stages using electrocardiographic (ECG) data. We have verified that GANSO can effectively improve the classifier performance, while the benchmark method SMOTE is not appropriate to deal with such a small size of the training set.

Sani, A., Lazaric, A., and Ryabko, D. (2015). “The replacement bootstrap for dependent data.” In: *IEEE International Symposium on Information Theory (ISIT)*. IEEE, pp. 1194–1198.

Applications that deal with time-series data often require evaluating complex statistics for which each time series is essentially one data point. When only a few time series are available, bootstrap methods are used to generate additional samples that can be used to evaluate empirically the statistic of interest. In this work a novel bootstrap method is proposed, which is shown to have some asymptotic consistency guarantees under the only assumption that the time series are stationary and ergodic. This contrasts previously available results that impose mixing or finite-memory assumptions on the data. Empirical evaluation on simulated and real data, using a practically relevant and complex extrema statistic is provided.

Sapp, T. R. A. (2017). “Efficient Estimation of Distributional Tail Shape and the Extremal Index with Applications to Risk Management.” In: *Journal of Mathematical Finance* 06(04), pp. 626–659.

Abundant evidence indicates that financial asset returns are thicker-tailed than a normal distribution would suggest. The most negative outcomes which carry the potential to wreak financial disaster also tend to be the most rare and may fall outside the scope of empirical observation. The difficulty of modelling these rare but extreme events has been greatly reduced by recent advances in extreme value theory (EVT). The tail shape parameter and the extremal index are the fundamental parameters governing the extreme behavior of the distribution, and the effectiveness of EVT in forecasting depends upon their reliable, accurate estimation. This study provides a comprehensive analysis of the performance of estimators of both key parameters. Five tail shape estimators are examined within a Monte Carlo setting along the dimensions of bias, variability, and probability prediction performance. Five estimators of the extremal index are also examined using Monte Carlo simulation. A recommended best estimator is selected in each case and applied within a Value at Risk context to the Wilshire 5000 index to illustrate its usefulness for risk measurement.

Sarris, D., Spiliotis, E., and Assimakopoulos, V. (2020). “Exploiting resampling techniques for model selection in forecasting: an empirical evaluation using out-of-sample tests.” In: *Operational Research* 20(2), pp. 701–721.

Model selection is a complex task widely examined in the literature due to the major gains in forecasting accuracy when performed successfully. To do so, many approaches have been proposed exploiting the available historical data in different ways. In-sample testing is the most common approach but is highly affected by the data and parameter estimation uncertainty. Out-of-sample tests, which use part of the data to evaluate the performance of the forecasting methods, are also well-known alternatives which usually lead to improvements. However, these are still vulnerable to data uncertainty such as noise and outliers. On the other hand, resampling techniques can be used to produce multiple clones of a time series with the same characteristics but a different component of randomness. In this paper, a model selection technique is proposed which takes advantage of the bootstrapping process to mitigate the effect of noise in the original data and then applies out-of-sample tests to the generated series to evaluate the forecasting performance of different methods. The approach is assessed across a large dataset of diverse time series and benchmarked versus other traditional approaches leading to promising results.

Saxena, D. and Cao, J. (2020). “Generative Adversarial Networks (GANs): Challenges, Solutions, and Future Directions.” In: *arXiv e-Print*.

Generative Adversarial Networks (GANs) is a novel class of deep generative models which has recently gained significant attention. GANs learn complex and high-dimensional distributions implicitly over images, audio, and data. However, there exist major challenges in training of GANs, i.e., mode collapse, non-convergence and instability, due to inappropriate design of network architecture, use of objective function and selection of optimization algorithm. Recently, to address these challenges, several solutions for better design and optimization of GANs have been investigated based on techniques of re-engineered network architectures, new objective functions and alternative optimization algorithms. To the best of our knowledge, there is no existing survey that has particularly focused on broad and systematic developments of these solutions. In this study, we perform a comprehensive survey of the advancements in GANs design and optimization solutions proposed to handle GANs challenges. We first identify key research issues within each design and optimization technique and then propose a new taxonomy to structure solutions by key research issues. In accordance with the taxonomy, we provide a detailed discussion on different GANs variants proposed within each solution and their relationships. Finally, based on the insights gained, we present the promising research directions in this rapidly growing field.

Saxena, D. and Cao, J. (2022). “Generative Adversarial Networks (GANs).” In: *ACM Computing Surveys* 54(3), pp. 1–42.

Generative Adversarial Networks (GANs) is a novel class of deep generative models that has recently gained significant attention. GANs learn complex and high-dimensional distributions implicitly over images, audio, and data. However, there exist major challenges in training of GANs, i.e., mode collapse, non-convergence, and instability, due to inappropriate design of network architecture, use of objective function, and selection of optimization algorithm. Recently, to address these challenges, several solutions for better design and optimization of GANs have been investigated based on techniques of re-engineered network architectures, new objective functions, and alternative optimization algorithms. To the best of our knowledge, there is no existing survey that has particularly focused on the broad and systematic developments of these solutions. In this study, we perform a comprehensive survey of the advancements in GANs design and optimization solutions proposed to handle GANs challenges. We first identify key research issues within each design and optimization technique and then propose a new taxonomy to structure solutions by key research issues. In accordance with the taxonomy, we provide a detailed discussion on different GANs variants proposed within each solution and their relationships. Finally, based on the insights gained, we present promising research directions in this rapidly growing field.

Schissler, A. G., Bedrick, E. J., Knudson, A. D., Kozubowski, T. J., Nguyen, T., Petereit, J., Piegorsch, W. W., and Tran, D. (2021). “Simulating High-Dimensional Multivariate Data using the bigsimr R Package.” In: *arXiv e-Print*.

It is critical to accurately simulate data when employing Monte Carlo techniques and evaluating statistical methodology. Measurements are often correlated and high dimensional in this era of big data, such as data obtained in high-throughput biomedical experiments. Due to the computational complexity and a lack of user-friendly software available to simulate these massive multivariate constructions, researchers resort to simulation designs that posit independence or perform arbitrary data transformations. To close this gap, we developed the Bigsimr Julia package with R and Python interfaces. This paper focuses on the R interface. These packages empower high-dimensional random vector simulation with arbitrary marginal distributions and dependency via a Pearson, Spearman, or Kendall correlation matrix. bigsimr contains high-performance features, including multi-core and graphical-processing-unit-accelerated algorithms to estimate correlation and compute the nearest correlation matrix. Monte Carlo studies quantify the accuracy and scalability of our approach, up to  $d = 10,000$ . We describe example workflows and apply to a high-dimensional data set – RNA-sequencing data obtained from breast cancer tumor samples.

Schneider, P. G. (2020). “Sparse economic scenarios.” In: *SSRN e-Print*.

We show how distributions can be reduced to low-dimensional scenario trees. Applied to intertemporal distributions, the scenarios and their probabilities become time-varying factors. From S&P 500 options, two or three time-varying scenarios suffice to forecast returns, implied variance or skewness on par, or better, than extant multivariate stochastic volatility jump-diffusion models, while reducing the computational effort to fractions of a second.

Schuderer, A., Bromuri, S., and Eekelen, M. van (2021). “Sim-Env: Decoupling OpenAI Gym Environments from Simulation Models.” In: *arXiv e-Print*.

Reinforcement learning (RL) is one of the most active fields of AI research. Despite the interest demonstrated by the research community in reinforcement learning, the development methodology still lags behind, with a severe lack of standard APIs to foster the development of RL applications. OpenAI Gym is probably the most



used environment to develop RL applications and simulations, but most of the abstractions proposed in such a framework are still assuming a semi-structured methodology. This is particularly relevant for agent-based models whose purpose is to analyse adaptive behaviour displayed by self-learning agents in the simulation. In order to bridge this gap, we present a workflow and tools for the decoupled development and maintenance of multi-purpose agent-based models and derived single-purpose reinforcement learning environments, enabling the researcher to swap out environments with ones representing different perspectives or different reward models, all while keeping the underlying domain model intact and separate. The Sim-Env Python library generates OpenAI-Gym-compatible reinforcement learning environments that use existing or purposely created domain models as their simulation back-ends. Its design emphasizes ease-of-use, modularity and code separation.

Schwaab, B., Zhang, X. (, and Lucas, A. (2021). “[Modeling Extreme Events: Time-Varying Extreme Tail Shape.](#)” In: *SSRN e-Print*.

A dynamic semi-parametric framework is proposed to study time variation in tail fatness of sovereign bond yield changes during the 2010–2012 euro area sovereign debt crisis measured at a high (15-minute) frequency. The framework builds on the Generalized Pareto Distribution (GPD) for modeling peaks over thresholds as in Extreme Value Theory, but casts the model in a conditional framework to allow for time-variation in the tail shape parameters. The score-driven updates used improve the expected Kullback-Leibler divergence between the model and the true data generating process on every step even if the GPD only fits approximately and the model is mis-specified, as will be the case in any finite sample. This is confirmed in simulations. Using the model, we find the ECB program had a beneficial impact on extreme upper tail quantiles, leaning against the risk of extremely adverse market outcomes while active.

Seabrook, I., Caccioli, F., and Aste, T. (2021). “[An Information Filtering approach to stress testing: an application to FTSE markets.](#)” In: *arXiv e-Print*.

We present a novel methodology to quantify the “impact” of and “response” to market shocks. We apply shocks to a group of stocks in a part of the market, and we quantify the effects in terms of average losses on another part of the market using a sparse probabilistic elliptical model for the multivariate return distribution of the whole market. Sparsity is introduced with an  $L_0$ -norm regularization, which forces to zero some elements of the inverse covariance according to a dependency structure inferred from an information filtering network. Our study concerns the FTSE 100 and 250 markets and analyzes impact and response to shocks both applied to and received from individual stocks and group of stocks. We observe that the shock pattern is related to the structure of the network associated with the sparse structure of the inverse covariance of stock returns. Central sectors appear more likely to be affected by shocks, and stocks with a large level of underlying diversification have a larger impact on the rest of the market when experiencing shocks. By analyzing the system during times of crisis and comparative market calmness, we observe changes in the shock patterns with a convergent behavior in times of crisis.

Sebastian, A. and Gebbie, T. (2019). “[Systematic Asset Allocation using Flexible Views for South African Markets.](#)” In: *arXiv e-Print*.

We implement a systematic asset allocation model using the Historical Simulation with Flexible Probabilities (HS-FP) framework developed by Meucci. The HS-FP framework is a flexible non-parametric estimation approach that considers future asset class behavior to be conditional on time and market environments, and derives a forward looking distribution that is consistent with this view while remaining close as possible to the prior distribution. The framework derives the forward looking distribution by applying unequal time and state conditioned probabilities to historical observations of asset class returns. This is achieved using relative entropy to find estimates with the least distortion to the prior distribution. Here, we use the HS-FP framework on South African financial market data for asset allocation purposes; by estimating expected returns, correlations and volatilities that are better represented through the measured market cycle. We demonstrated a range of state variables that can be useful towards understanding market environments. Concretely, we compare the out-of-sample performance for a specific configuration of the HS-FP model relative to classic Mean Variance Optimization(MVO) and Equally Weighted (EW) benchmark models. The framework displays low probability of backtest overfitting and the out-of-sample net returns and Sharpe ratio point estimates of the HS-FP model outperforms the benchmark models. However, the results are inconsistent when training windows are varied, the Sharpe ratio is seen to be inflated, and the method does not demonstrate statistically significant out-performance on a gross and net basis.

Shah, V. and Shroff, G. (2021). “[Forecasting Market Prices using DL with Data Augmentation and Meta-learning: ARIMA still wins!](#)” In: *arXiv e-Print*.

Deep-learning techniques have been successfully used for time-series forecasting and have often shown superior performance on many standard benchmark datasets as compared to traditional techniques. Here we present a comprehensive and comparative study of performance of deep-learning techniques for forecasting prices in financial markets. We benchmark state-of-the-art deep-learning baselines, such as NBeats, etc., on data from currency as well as stock markets. We also generate synthetic data using a fuzzy-logic based model of demand driven by technical rules such as moving averages, which are often used by traders. We benchmark the baseline techniques on this synthetic data as well as use it for data augmentation. We also apply gradient-based meta-learning to account for non-stationarity of financial time-series. Our extensive experiments notwithstanding, the surprising result is that the standard ARIMA models outperforms deep-learning even using data augmentation or meta-learning. We conclude by speculating as to why this might be the case.

Sharma, D., Bouchaud, J.-P., Gualdi, S., Tarzia, M., and Zamponi, F. (2021). “V- U-, L- or W- shaped economic recovery after Covid-19: Insights from an Agent Based Model.” In: *PLOS ONE* 16(3). Ed. by S. C. Gherghina, e0247823.

We discuss the impact of a Covid-19-like shock on a simple model economy, described by the previously developed Mark-0 Agent-Based Model. We consider a mixed supply and demand shock, and show that depending on the shock parameters (amplitude and duration), our model economy can display V-shaped, U-shaped or W-shaped recoveries, and even an L-shaped output curve with permanent output loss. This is due to the economy getting trapped in a self-sustained “bad” state. We then discuss two policies that attempt to moderate the impact of the shock: giving easy credit to firms, and the so-called helicopter money, i.e. injecting new money into the households savings. We find that both policies are effective if strong enough. We highlight the potential danger of terminating these policies too early, although inflation is substantially increased by lax access to credit. Finally, we consider the impact of a second lockdown. While we only discuss a limited number of scenarios, our model is flexible and versatile enough to accommodate a wide variety of situations, thus serving as a useful exploratory tool for a qualitative, scenario-based understanding of post-Covid recovery. The corresponding code is available on-line.

Sharma, D., Bouchaud, J.-P., Tarzia, M., and Zamponi, F. (2020). “A constraint-satisfaction agent-based model for the macroeconomy.” In: *arXiv e-Print*.

We introduce a prototype agent-based model of the macroeconomy, with a budgetary constraint at its core. The model is related to a class of constraint satisfaction problems, which has been thoroughly investigated in computer science. We identify three different regimes of our toy economy upon varying the amount of debt that each agent can accumulate before defaulting. In presence of a very loose constraint on debt, endogenous crises leading to waves of synchronized bankruptcies are present. In the opposite regime of very tight debt constraining, the bankruptcy rate is extremely high and the economy remains structure-less. In an intermediate regime, the economy is stable with very low bankruptcy rate and no aggregate-level crises. This third regime displays a rich phenomenology: the system spontaneously and dynamically self-organizes in a set of cheap and expensive goods (i.e. some kind of “speciation”), with switches triggered by random fluctuations and feedback loops. Our analysis confirms the central role that debt levels play in the stability of the economy.

Shen, Z., Liu, J., He, Y., Zhang, X., Xu, R., Yu, H., and Cui, P. (2021). “Towards Out-Of-Distribution Generalization: A Survey.” In: *arXiv e-Print*.

Classic machine learning methods are built on the *i.i.d.* assumption that training and testing data are independent and identically distributed. However, in real scenarios, the *i.i.d.* assumption can hardly be satisfied, rendering the sharp drop of classic machine learning algorithms’ performances under distributional shifts, which indicates the significance of investigating the Out-of-Distribution generalization problem. Out-of-Distribution (OOD) generalization problem addresses the challenging setting where the testing distribution is unknown and different from the training. This paper serves as the first effort to systematically and comprehensively discuss the OOD generalization problem, from the definition, methodology, evaluation to the implications and future directions. Firstly, we provide the formal definition of the OOD generalization problem. Secondly, existing methods are categorized into three parts based on their positions in the whole learning pipeline, namely unsupervised representation learning, supervised model learning and optimization, and typical methods for each category are discussed in detail. We then demonstrate the theoretical connections of different categories, and introduce the commonly used datasets and evaluation metrics. Finally, we summarize the whole literature and raise some future directions for OOD generalization problem. The summary of OOD generalization methods reviewed in this survey can be found at <http://out-of-distribution-generalization.com>.

Shi, Y., Li, B., and Du, G. (2021). “Pyramid scheme in stock market: a kind of financial market simulation.” In: *arXiv e-Print*.

Artificial stock market simulation based on agent is an important means to study financial market. Based on the assumption that the investors are composed of a main fund, small trend and contrarian investors characterized by four parameters, we simulate and research a kind of financial phenomenon with the characteristics of pyramid schemes. Our simulation results and theoretical analysis reveal the relationships between the rate of return of the main fund and the proportion of the trend investors in all small investors, the small investors’ parameters of taking profit and stopping loss, the order size of the main fund and the strategies adopted by the main fund. Our work are helpful to explain the financial phenomenon with the characteristics of pyramid schemes in financial markets, design trading rules for regulators and develop trading strategies for investors.

Siddique, A., Hasan, I., and Lynch, D. (2019). *Stress Testing (Second Edition)*. Risk Books. 500 pp.

A financial institution’s repertoire of tools has to be broad and dynamic in the post-crisis era. Stress testing has come a long way since the first edition, thinking has changed dramatically, and so the second edition of *Stress Testing: Approaches, Methods and Applications* has added chapters that address these refinements in thinking and deals with new topics such as pre-position net revenue. The authoritative guide and reference tool for stress testing, this book is essential for risk managers, regulators and consultants who want a clearer understanding of the methods, tools and uses of stress testing in different risk areas. Described as an industry in itself, stress testing constitutes an extensive element of financial institutions risk management and capital-adequacy assessments. Definitions of what constitutes stress testing and how stress testing should be used had been hard to pin down meaning it can be hard to benchmark one’s own firm in this area. There is now a dearth of published information on stress testing approaches and it can be a time-consuming process working out which approach is best for your firm, often resulting in banks hiring expensive management consultants. The aim of this book is to help CROs, CFOs or those working in the treasury space to figure out which approach will be most appropriate. Offering insights and guidelines that expound the various approaches and highlights those most appropriate with regard to the guidance.

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.

Siew, L. W., Hj, and Ismail, H. B. (2015). “The Impact of Different Economic Scenarios Towards Portfolio Selection in Enhanced Index Tracking Problem.” In: *Advanced Science Letters*, pp. 1285–1288.

Enhanced index tracking is a portfolio management which aims to track and outperform the benchmark stock market index without purchasing all stocks from the benchmark index components. Enhanced index tracking is a portfolio selection problem which can be represented by an optimization model. The objective of this paper is to study the impact of three different economic scenarios towards portfolio selection and portfolio performance in enhanced index tracking problem. The economic scenarios are divided into three sub-periods representing the growth period in the economy, financial crisis and the recovery period. In this study, the optimization model with sum weighted approach is applied in constructing the optimal portfolio. This paper also studies the consistency of the optimal portfolios to outperform the benchmark index for the three economic scenarios. The data consists of weekly price of 24 components stocks in Malaysia main market index which is Kuala Lumpur Composite Index from January 1994 until June 2008. The results of this study show different optimal portfolio selections and performances for the three economic scenarios. Besides that, the optimal portfolios are able to generate higher mean return than the benchmark index return with only selecting 10 percent from the benchmark index components for the three economic scenarios. Therefore, the optimization model with sum weighted approach

is appropriate for the investors in Malaysia. The significance of this study is to identify the optimal portfolios for three different economic scenarios to track and outperform Malaysia market index consistently. In addition, this study concludes that the performance of the optimal portfolio is the highest during economic recovery in Malaysia.

Silva, A. C. and Ferreira, F. F. (2021). “Surrogate Monte Carlo.” In: *arXiv e-Print*.

This article proposes an artificial data generating algorithm that is simple and easy to customize. The fundamental concept is to perform random permutation of Monte Carlo generated random numbers which conform to the unconditional probability distribution of the original real time series. Similar to constraint surrogate methods, random permutations are only accepted if a given objective function is minimized. The objective function is selected in order to describe the most important features of the stochastic process. The algorithm is demonstrated by producing simulated log-returns of the S&P 500 stock index.

Silva, T., Pinheiro, P. R., and Poggi, M. (2017). “A more human-like portfolio optimization approach.” In: *European Journal of Operational Research* 256(1), pp. 252–260.

We propose a new method for constructing a portfolio aligned with the investor’s profile and personal perspective. We propose a new way to create Black-Litterman views. Our method is more qualitative than other Black-Litterman variants and helps to quantify the subjective views. We analyze the sensitivity of the method through practical experiments. Black and Litterman proposed an improvement to the Markowitz portfolio optimization model. They suggested the construction of views to represent investor’s opinion about the future of stocks’ returns. However, conceiving these views can be quite confusing. It requires the investor to quantify several subjective parameters. In this article, we propose a new way of creating these views using Verbal Decision Analysis. Questionnaires were designed with the intent of making it easier for investors to express their vision about stocks. Following the ZAPROS methodology, the investor answers sets of questions allowing to determine a Formal Index of Quality (FIQ). The views are then derived from the resulting FIQ. Our approach was implemented and tested on data from the Brazilian Stocks. It allows investors to create a personal risk-return balanced portfolio without the help of an expert. The experiments show that the proposed method mitigates the impact of poor view estimation. Also, one must notice that the method is qualitative and its aim is to create a more efficient portfolio considering the investor’s vision.

Skoglund, J. (2019). “Quantification of model risk in stress testing and scenario analysis.” In: *Journal of Risk Model Validation* 13(1), pp. 1–23.

Being able to understand and quantify the model risk inherent in loss-projection models used in macroeconomic stress testing and impairment estimation is a significant concern for both banks and regulators. The application of relative entropy techniques allows model misspecification robustness to be numerically quantified using exponential tilting toward an alternative probability law. Employing a particular loss-forecasting model, we quantify the worst-case-loss term structures of that model, yielding insights into the behavior of the worst-case scenario. In general, the worst case obtained represents an upward scaling of the term structure consistent with the exponential tilting adjustment. The relative entropy approach to model risk we use has its foundation in economics with robust forecasting analysis, and it has recently started to be applied in risk management. This technique can complement traditional model risk quantification techniques, where a specific direction or range of model misspecification reasons are usually considered, such as model sensitivity analysis, model parameter uncertainty analysis, competing models and conservative model assumptions.

Smith, K. E. and Smith, A. O. (2020). “Conditional GAN for timeseries generation.” In: *arXiv e-Print*.

It is abundantly clear that time dependent data is a vital source of information in the world. The challenge has been for applications in machine learning to gain access to a considerable amount of quality data needed for algorithm development and analysis. Modeling synthetic data using a Generative Adversarial Network (GAN) has been at the heart of providing a viable solution. Our work focuses on one dimensional times series and explores the few shot approach, which is the ability of an algorithm to perform well with limited data. This work attempts to ease the frustration by proposing a new architecture, Time Series GAN (TSGAN), to model realistic time series data. We evaluate TSGAN on 70 data sets from a benchmark time series database. Our results demonstrate that TSGAN performs better than the competition both quantitatively using the Frechet Inception Score (FID) metric, and qualitatively when classification is used as the evaluation criteria.

Snow, D. (2020a). “DataGene: A Framework for Dataset Similarity.” In: *SSRN e-Print*.

DataGene is developed to identify data set similarity between real and synthetic datasets as well as train, test, and validation datasets. For many modelling and software development tasks there is a need for datasets to have share similar characteristics. This has traditionally been achieved with visualizations, DataGene seeks to

replace these visual methods with a range of novel quantitative methods. Please see the GitHub repository to inspect and install the Python code.

Snow, D. (2020b). “DeltaPy: A Framework for Tabular Data Augmentation in Python.” In: *SSRN e-Print*.

A range of data abstractions have come to the fore since the re-emergence of machine learning. This includes procedures like feature engineering, extraction, transformation, and selection, as well as data preprocessing, generation, synthesis, and augmentation. This report attempts to unify some of this terminology with the development of a bare-bones Python package, DeltaPy.

Staum, J. (2009). “Monte Carlo Computation in Finance.” In: *Monte Carlo and Quasi-Monte Carlo Methods 2008*. Ed. by P. L’Ecuyer and A. B. Owen. Springer Berlin Heidelberg, pp. 19–42.

This advanced tutorial aims at an exposition of problems in finance that are worthy of study by the Monte Carlo research community. It describes problems in valuing and hedging securities, risk management, portfolio optimization, and model calibration. It surveys some areas of active research in efficient procedures for simulation in finance and addresses the impact of the business context on the opportunities for efficiency. There is an emphasis on the many challenging problems in which it is necessary to perform several similar simulations.

Steehouwer, H. (2014). “Ex-ante risk management with scenarios.” In: *CFA Society Switzerland Series*.

Ex-ante risk management with help of scenarios including important information such as the impact of the business cycle is an innovative methodology used by sophisticated managers gaining worldwide popularity. Advanced stochastic scenario analysis supports financial decision-making when dealing with uncertainty through accounting for structural changes impacting trend values as well as short term events influencing the market momentum. The extent to which investor objectives will be achieved without violating risk limits is influenced by the path that financial and economic risk drivers will follow. A frequency domain based approach differs from traditional simulation approaches through the creation of plausible and relevant scenarios. The Dynamic Scenario Generator (DSG) employs a unique combination of statistical and econometric techniques. The methodology is a mixture of frequency domain techniques, dynamic factor models and special approaches for dealing with non-normal distribution characteristics such as skewness and fat tails. The main purpose of the central frequency domain methodology is to describe all empirical features (stylized facts) of the time series behavior of financial and economic variables simultaneously, rather than focusing on only one or a few aspects in isolation.

Steehouwer, H. and Slater, A. (2010). “Macroeconomic Scenarios: A Frequency Domain Approach.” In: *SSRN e-Print*.

Here we describe and motivate a central frequency domain methodology for time series modeling as proposed in Steehouwer (2005) which consists of the following steps

- Time series decomposition
- Time series analysis
- Model specification and estimation
- Model analysis (not included)

Stephanovitch, A., Tanielian, U., Cadre, B., Klutchnikoff, N., and Biau, G. (2022). “Optimal 1-Wasserstein Distance for WGANs.” In: *arXiv e-Print*.

The mathematical forces at work behind Generative Adversarial Networks raise challenging theoretical issues. Motivated by the important question of characterizing the geometrical properties of the generated distributions, we provide a thorough analysis of Wasserstein GANs (WGANs) in both the finite sample and asymptotic regimes. We study the specific case where the latent space is univariate and derive results valid regardless of the dimension of the output space. We show in particular that for a fixed sample size, the optimal WGANs are closely linked with connected paths minimizing the sum of the squared Euclidean distances between the sample points. We also highlight the fact that WGANs are able to approach (for the 1-Wasserstein distance) the target distribution as the sample size tends to infinity, at a given convergence rate and provided the family of generative Lipschitz functions grows appropriately. We derive in passing new results on optimal transport theory in the semi-discrete setting.

Sun, H., Deng, Z., Chen, H., and Parkes, D. C. (2020). “Decision-Aware Conditional GANs for Time Series Data.” In: *arXiv e-Print*.



We introduce the decision-aware time-series conditional generative adversarial network (DAT-CGAN) as a method for time-series generation. The framework adopts a multi-Wasserstein loss on structured decision-related quantities, capturing the heterogeneity of decision-related data and providing new effectiveness in supporting the decision processes of end users. We improve sample efficiency through an overlapped block-sampling method, and provide a theoretical characterization of the generalization properties of DAT-CGAN. The framework is demonstrated on financial time series for a multi-time-step portfolio choice problem. We demonstrate better generative quality in regard to underlying data and different decision-related quantities than strong, GAN-based baselines.

Tadayon, M. and Pottie, G. (2020). “tsBNgen: A Python Library to Generate Time Series Data from an Arbitrary Dynamic Bayesian Network Structure.” In: *arXiv e-Print*.

Synthetic data is widely used in various domains. This is because many modern algorithms require lots of data for efficient training, and data collection and labeling usually are a time-consuming process and are prone to errors. Furthermore, some real-world data, due to its nature, is confidential and cannot be shared. Bayesian networks are a type of probabilistic graphical model widely used to model the uncertainties in real-world processes. Dynamic Bayesian networks are a special class of Bayesian networks that model temporal and time series data. In this paper, we introduce the tsBNgen, a Python library to generate time series and sequential data based on an arbitrary dynamic Bayesian network. The package, documentation, and examples can be downloaded from <https://github.com/manitadayon/tsBNgen>.

Takahashi, S., Chen, Y., and Tanaka-Ishii, K. (2019). “Modeling financial time-series with generative adversarial networks.” In: *Physica A: Statistical Mechanics and its Applications* 527, p. 121261.

Abstract Financial time-series modeling is a challenging problem as it retains various complex statistical properties and the mechanism behind the process is unrevealed to a large extent. In this paper, a deep neural networks based approach, generative adversarial networks (GANs) for financial time-series modeling is presented. GANs learn the properties of data and generate realistic data in a data-driven manner. The GAN model produces a time-series that recovers the statistical properties of financial time-series such as the linear unpredictability, the heavy-tailed price return distribution, volatility clustering, leverage effects, the coarse-fine volatility correlation, and the gain/loss asymmetry.

Tan, V. and Zohren, S. (2021). “Large Non-Stationary Noisy Covariance Matrices: A Cross-Validation Approach.” In: *SSRN e-Print*.

We introduce a novel covariance estimator that exploits the heteroskedastic nature of financial time series by employing exponential weighted moving averages and shrinking the in-sample eigenvalues through cross-validation. Our estimator is model-agnostic in that we make no assumptions on the distribution of the random entries of the matrix or structure of the covariance matrix. Additionally, we show how Random Matrix Theory can provide guidance for automatic tuning of the hyperparameter which characterizes the time scale for the dynamics of the estimator. By attenuating the noise from both the cross-sectional and time-series dimensions, we empirically demonstrate the superiority of our estimator over competing estimators that are based on exponentially-weighted and uniformly-weighted covariance matrices.

Tanaka, K. (2017). “Forecasting scenarios from the perspective of a reverse stress test using second-order cone programming.” In: *Journal of Risk Model Validation* 11(2), pp. 1–21.

This paper proposes a model for forecasting scenarios from the perspective of a reverse stress test using interest rate (JGB10Y yield), equity (Nikkei 225) and foreign exchange (USD/U) data. The model consists of a constraint with error terms of dynamic conditional correlation-generalized autoregressive conditional heteroscedasticity (DCC-GARCH) for expressing risk factors (RFs) located in an acceptable range, where the acceptable range is determined by the Mahalanobis distance, which consists of error terms of DCC-GARCH and the correlation between one RF and another RF; maximization of the objection function, which is the loss of market portfolio (ie, minimization of the difference of a market portfolio). I also show that forecasting scenarios identified by this model are valid in terms of expressing very stressful data, which documents that some financial institutions may be in default, and that there is a mostly distributed multivariate normal distribution; and this model can be solved by formulating second-order cone programming, which is standard in the field of mathematical optimization programming.

Taylor, L. and Nitschke, G. (2017). “Improving Deep Learning using Generic Data Augmentation.” In: *arXiv e-Print*.

Deep artificial neural networks require a large corpus of training data in order to effectively learn, where collection of such training data is often expensive and laborious. Data augmentation overcomes this issue by artificially inflating the training set with label preserving transformations. Recently there has been extensive use of generic data augmentation to improve Convolutional Neural Network (CNN) task performance. This study benchmarks

various popular data augmentation schemes to allow researchers to make informed decisions as to which training methods are most appropriate for their data sets. Various geometric and photometric schemes are evaluated on a coarse-grained data set using a relatively simple CNN. Experimental results, run using 4-fold cross-validation and reported in terms of Top-1 and Top-5 accuracy, indicate that cropping in geometric augmentation significantly increases CNN task performance.

ter Ellen, S., Hommes, C. H., and Zwinkels, R. C. J. (2021). “Comparing behavioural heterogeneity across asset classes.” In: *Journal of Economic Behavior & Organization*.

We estimate an endowment-based asset pricing model in which agents have heterogeneous and time-varying beliefs about the future price on a range of asset classes. This gives insight into the extent behaviour differs across assets, and what this implies for market stability. We find evidence for behavioural heterogeneity for all asset classes but equity. Heterogeneity is especially large and persistent in asset classes for which limits to arbitrage are more binding. In less constrained (financial) markets, agents update their beliefs more frequently. Consequently, the probability of behavioural bubbles and crashes is substantially higher in macroeconomic asset classes than in financial asset classes.

Thiem, T. N., Kemeth, F. P., Bertalan, T., Laing, C. R., and Kevrekidis, I. G. (2021). “Global and Local Reduced Models for Interacting, Heterogeneous Agents.” In: *arXiv e-Print*.

Large collections of coupled, heterogeneous agents can manifest complex dynamical behavior presenting difficulties for simulation and analysis. However, if the collective dynamics lie on a low-dimensional manifold then the original agent-based model may be approximated with a simplified surrogate model on and near the low-dimensional space where the dynamics live. This is typically accomplished by deriving coarse variables that summarize the collective dynamics, these may take the form of either a collection of scalars or continuous fields (e.g. densities), which are then used as part of a reduced model. Analytically identifying such simplified models is challenging and has traditionally been accomplished through the use of mean-field reductions or an Ott-Antonsen ansatz, but is often impossible. Here we present a data-driven coarse-graining methodology for discovering such reduced models. We consider two types of reduced models: globally-based models which use global information and predict dynamics using information from the whole ensemble, and locally-based models that use local information, that is, information from just a subset of agents close (close in heterogeneity space, not physical space) to an agent, to predict the dynamics of an agent. For both approaches we are able to learn laws governing the behavior of the reduced system on the low-dimensional manifold directly from time series of states from the agent-based system. A nontrivial conclusion is that the dynamics can be equally well reproduced by an all-to-all coupled as well as by a locally coupled model of the same agents.

Traccucci, P., Dumontier, L., Garchery, G., and Jacot, B. (2019). “A Triptych Approach for Reverse Stress Testing of Complex Portfolios.” In: *Risk (Cutting Edge)*.

Pascal Traccucci, Luc Dumontier, Guillaume Garchery and Benjamin Jacot present an extended reverse stress test (ERST) triptych approach with three variables: level of plausibility, level of loss and scenario. Any two of these variables can be derived, provided the third is given as input. A new version of the Levenberg-Marquardt optimisation algorithm is introduced to derive the ERST in certain complex cases.

Tran, T., Pham, T., Carneiro, G., Palmer, L., and Reid, I. (2017). “A Bayesian Data Augmentation Approach for Learning Deep Models.” In: *arXiv e-Print*.

Data augmentation is an essential part of the training process applied to deep learning models. The motivation is that a robust training process for deep learning models depends on large annotated datasets, which are expensive to be acquired, stored and processed. Therefore a reasonable alternative is to be able to automatically generate new annotated training samples using a process known as data augmentation. The dominant data augmentation approach in the field assumes that new training samples can be obtained via random geometric or appearance transformations applied to annotated training samples, but this is a strong assumption because it is unclear if this is a reliable generative model for producing new training samples. In this paper, we provide a novel Bayesian formulation to data augmentation, where new annotated training points are treated as missing variables and generated based on the distribution learned from the training set. For learning, we introduce a theoretically sound algorithm — generalised Monte Carlo expectation maximisation, and demonstrate one possible implementation via an extension of the Generative Adversarial Network (GAN). Classification results on MNIST, CIFAR-10 and CIFAR-100 show the better performance of our proposed method compared to the current dominant data augmentation approach mentioned above — the results also show that our approach produces better classification results than similar GAN models.

Trimborn, T., Otte, P., Cramer, S., Beikirch, M., Pabich, E., and Frank, M. (2020). “SABCEMM: A Simulator for Agent-Based Computational Economic Market Models.” In: *Computational Economics* 55(2), pp. 707–744.

We introduce the simulation tool SABCEMM (Simulator for Agent-Based Computational Economic Market Models) for agent-based computational economic market (ABCEM) models. Our simulation tool is implemented in C++ and we can easily run ABCEM models with several million agents. The object-oriented software design enables the isolated implementation of building blocks for ABCEM models, such as agent types and market mechanisms. The user can design and compare ABCEM models in a unified environment by recombining existing building blocks using the XML-based SABCEMM configuration file. We introduce an abstract ABCEM model class which our simulation tool is built upon. Furthermore, we present the software architecture as well as computational aspects of SABCEMM. Here, we focus on the efficiency of SABCEMM with respect to the run time of our simulations. We show the great impact of different random number generators on the run time of ABCEM models. The code and documentation is published on GitHub at <https://github.com/SABCEMM/SABCEMM>, such that all results can be reproduced by the reader.

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.

Tsamardinos, I., Greasidou, E., and Borboudakis, G. (2018). “Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation.” In: *Machine Learning*, pp. 1–28.

Cross-Validation (CV), and out-of-sample performance-estimation protocols in general, are often employed both for (a) selecting the optimal combination of algorithms and values of hyper-parameters (called a configuration) for producing the final predictive model, and (b) estimating the predictive performance of the final model. However, the cross-validated performance of the best configuration is optimistically biased. We present an efficient bootstrap method that corrects for the bias, called Bootstrap Bias Corrected CV (BBC-CV). BBC-CV main idea is to bootstrap the whole process of selecting the best-performing configuration on the out-of-sample predictions of each configuration, without additional training of models. In comparison to the alternatives, namely the nested cross-validation (Varma and Simon in BMC Bioinform 7(1):91, 2006) and a method by Tibshirani and Tibshirani (Ann Appl Stat 822-829, 2009), BBC-CV is computationally more efficient, has smaller variance and bias, and is applicable to any metric of performance (accuracy, AUC, concordance index, mean squared error). Subsequently, we employ again the idea of bootstrapping the out-of-sample predictions to speed up the CV process. Specifically, using a bootstrap-based statistical criterion we stop training of models on new folds of inferior (with high probability) configurations. We name the method Bootstrap Bias Corrected with Dropping CV (BBCD-CV) that is both efficient and provides accurate performance estimates.

Tsanakas, A. and Zhu, R. (2021). “Copula model selection using image recognition.” In: *SSRN e-Print*.

The choice of a copula model from limited data is a hard but important task. Motivated by the visual patterns that different copula models produce in smoothed density heatmaps, we consider copula model selection as an image recognition problem. We extract image features from heatmaps using the pre-trained AlexNet, and present workflows for model selection that combine image features with statistical information. We employ dimension reduction via Principal Component and Linear Discriminant Analyses, and use a Support Vector Machine classifier. Simulation studies show that the use of image data improves the accuracy of the copula model selection task, particularly in scenarios where sample sizes and correlations are low. This finding indicates that transfer learning can support statistical procedures of model selection.

van Beek, M. (2020). “Consistent Calibration of Economic Scenario Generators: The Case for Conditional Simulation.” In: *arXiv e-Print*.

Economic Scenario Generators (ESGs) simulate economic and financial variables forward in time for risk management and asset allocation purposes. It is often not feasible to calibrate the dynamics of all variables within the ESG to historical data alone. Calibration to forward-information such as future scenarios and return expectations is needed for stress testing and portfolio optimization, but no generally accepted methodology is available.

This paper introduces the Conditional Scenario Simulator, which is a framework for consistently calibrating simulations and projections of economic and financial variables both to historical data and forward-looking information. The framework can be viewed as a multi-period, multi-factor generalization of the Black-Litterman model, and can embed a wide array of financial and macroeconomic models. Two practical examples demonstrate this in a frequentist and Bayesian setting.

Vandin, A., Giachini, D., Lamperti, F., and Chiaromonte, F. (2021). “Automated and Distributed Statistical Analysis of Economic Agent-Based Models.” In: *arXiv e-Print*.

We propose a novel approach to the statistical analysis of simulation models and, especially, agent-based models (ABMs). Our main goal is to provide a fully automated and model-independent tool-kit to inspect simulations and perform counterfactual analysis. Our approach: (i) is easy-to-use by the modeller, (ii) improves reproducibility of results, (iii) optimizes running time given the modeller’s machine, (iv) automatically chooses the number of required simulations and simulation steps to reach user-specified statistical confidence, and (v) automatically performs a variety of statistical tests. In particular, our framework is designed to distinguish the transient dynamics of the model from its steady-state behaviour (if any), estimate properties of the model in both “phases”, and provide indications on the ergodic (or non-ergodic) nature of the simulated processes – which, in turns allows one to gauge the reliability of a steady-state analysis. Estimates are equipped with statistical guarantees, allowing for robust comparisons across computational experiments. To demonstrate the effectiveness of our approach, we apply it to two models from the literature: a large scale macro-financial ABM and a small scale prediction market model. Compared to prior analyses of these models, we obtain new insights and we are able to identify and fix some erroneous conclusions.

Vieira, A. (2020). “Generating Synthetic Sequential Data using GANs.” In: *Towards AI*.

In this post, we describe and apply an extended version of a recent powerful method to generate synthetic sequential data - DoppelGANger. It is a framework based on Generative Adversarial Networks (GANs) with some innovations that make possible the generation of synthetic versions of complex sequential datasets.

Vinod, H. D. (2012). “Constructing Scenarios of Time Heterogeneous Series for Stress Testing.” In: *SSRN e-Print*.

Heterogeneous global trends in asset prices and savings affect the macro economy. Our challenge is to use limited data to make inference regarding underlying causes. In general, government and business decision makers, FDIC type regulators and risk professionals need quantitative tools to help generate plausible scenarios of state-dependent and time heterogeneous nonstationary time series. We suggest using maximum entropy type bootstraps, recently implemented in an R software package called “meboot.” A new modification of meboot divides the data series into blocks and can randomly modify the (down, at or up) direction of series within each block. Our large number of resamples are then available for construction of scenarios for probabilistic stress testing. A simulation study evaluates the performance of our proposal in the context of many types of time-heterogeneity showing that it behaves better than moving block bootstraps. We apply meboot tools to stress test inference regarding Granger-causality between asset prices and world savings rates, and also to the ‘Value at Risk’ used in Finance.

Vinod, H. D. (2021). “R Package GeneralCorr Functions for Portfolio Choice.” In: *SSRN e-Print*.

We explain the usage of the following new R functions in my package called ‘generalCorr’ Vinod (2021b). The function `sudoCoeffParcor()` is for pseudo regression coefficients from kernel regressions. They are a nonlinear version of regression coefficients from standardized data. The functions `decileVote()`, `momentVote()`, `exactSdMtx()` compute exact computation of stochastic dominance from ECDF areas. `dif4mtx()` computes growth, change in growth, etc., up-to order 4 differencing of time series. The function `bootDom12()` reports bootstrap confidence intervals for any confidence level. We illustrate all these functions with a toy example of only seven observations. The last section has the code which produced various tables in this document.

Volpi, R., Namkoong, H., Sener, O., Duchi, J., Murino, V., and Savarese, S. (2018). “Generalizing to Unseen Domains via Adversarial Data Augmentation.” In: *arXiv e-Print*.

We are concerned with learning models that generalize well to different unseen domains. We consider a worst-case formulation over data distributions that are near the source domain in the feature space. Only using training data from a single source distribution, we propose an iterative procedure that augments the dataset with examples from a fictitious target domain that is “hard” under the current model. We show that our iterative scheme is an adaptive data augmentation method where we append adversarial examples at each iteration. For softmax losses, we show that our method is a data-dependent regularization scheme that behaves differently from classical regularizers that regularize towards zero (e.g., ridge or lasso). On digit recognition and semantic segmentation tasks, our method learns models improve performance across a range of a priori unknown target domains.

Waller, N. G. (2020). “Generating Correlation Matrices With Specified Eigenvalues Using the Method of Alternating Projections.” In: *The American Statistician* 74(1), pp. 21–28.

This article describes a new algorithm for generating correlation matrices with specified eigenvalues. The algorithm uses the method of alternating projections (MAP) that was first described by Neumann. The MAP algorithm for generating correlation matrices is both easy to understand and to program in higher-level computer languages, making this method accessible to applied researchers with no formal training in advanced mathematics. Simulations indicate that the new algorithm has excellent convergence properties. Correlation matrices with specified eigenvalues can be profitably used in Monte Carlo research in statistics, psychometrics, computer science, and related disciplines. To encourage such use, R code (R Core Team) for implementing the algorithm is provided in the supplementary material.

Wang, B., Xu, H., Zhang, J., Chen, C., Fang, X., Xu, Y., Kang, N., Hong, L., Jiang, C., Cai, X., Li, J., Zhou, F., Li, Y., Liu, Z., Chen, X., Han, K., Shu, H., Song, D., Wang, Y., Zhang, W., Xu, C., Li, Z., Liu, W., and Zhang, T. (2020a). “VEGA: Towards an End-to-End Configurable AutoML Pipeline.” In: *arXiv e-Print*.

Automated Machine Learning (AutoML) is an important industrial solution for automatic discovery and deployment of the machine learning models. However, designing an integrated AutoML system faces four great challenges of configurability, scalability, integrability, and platform diversity. In this work, we present VEGA, an efficient and comprehensive AutoML framework that is compatible and optimized for multiple hardware platforms. a) The VEGA pipeline integrates various modules of AutoML, including Neural Architecture Search (NAS), Hyperparameter Optimization (HPO), Auto Data Augmentation, Model Compression, and Fully Train. b) To support a variety of search algorithms and tasks, we design a novel fine-grained search space and its description language to enable easy adaptation to different search algorithms and tasks. c) We abstract the common components of deep learning frameworks into a unified interface. VEGA can be executed with multiple back-ends and hardware. Extensive benchmark experiments on multiple tasks demonstrate that VEGA can improve the existing AutoML algorithms and discover new high-performance models against SOTA methods, e.g. the searched DNet model zoo for Ascend 10x faster than EfficientNet-B5 and 9.2x faster than RegNetX-32GF on ImageNet. VEGA is open-sourced at <https://github.com/huawei-noah/vega>.

Wang, H. and Tu, W. (2014). “Bootstrap methods: the classical theory and recent development.” In: *Wiley statsref: statistics reference online*. Ed. by N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, and J. L. Teugels. Chichester, UK: John Wiley & Sons, Ltd, pp. 1–12.

Bootstrap is a resampling method for statistical inference. Under fairly general conditions, the technique can be used to approximate sampling distributions of almost any statistics, by recycling data from the observed sample, that is, resampling. In this article, we review the theoretical tenets of bootstrapping, focusing primarily on the fundamental property of consistency, while showing examples where lack of consistency can lead to failures of the method. We also describe residual and pairs bootstrap methods in linear models, as well as their applications in low- and high-dimensional problems. Finally, we discuss a modified bootstrap procedure in big data situations.

Wang, L., Ma, F., and Liu, G. (2020b). “Forecasting stock volatility in the presence of extreme shocks: Short-term and long-term effects.” In: *Journal of Forecasting* 39(5), pp. 797–810.

This paper introduces a novel generalized autoregressive conditional heteroskedasticity-mixed data sampling-extreme shocks (GARCH-MIDAS-ES) model for stock volatility to examine whether the importance of extreme shocks changes in different time ranges. Based on different combinations of the short- and long-term effects caused by extreme events, we extend the standard GARCH-MIDAS model to characterize the different responses of the stock market for short- and long-term horizons, separately or in combination. The unique timespan of nearly 100 years of the Dow Jones Industrial Average (DJIA) daily returns allows us to understand the stock market volatility under extreme shocks from a historical perspective. The in-sample empirical results clearly show that the DJIA stock volatility is best fitted to the GARCH-MIDAS-SLES model by including the short- and long-term impacts of extreme shocks for all forecasting horizons. The out-of-sample results and robustness tests emphasize the significance of decomposing the effect of extreme shocks into short- and long-term effects to improve the accuracy of the DJIA volatility forecasts.

Wang, R. (2021). “Discriminating modelling approaches for Point in Time Economic Scenario Generation.” In: *arXiv e-Print*.

We introduce the notion of Point in Time Economic Scenario Generation (PiT ESG) with a clear mathematical problem formulation to unify and compare economic scenario generation approaches conditional on forward looking market data. Such PiT ESGs should provide quicker and more flexible reactions to sudden economic changes than traditional ESGs calibrated solely to long periods of historical data. We specifically take as economic



variable the S&P500 Index with the VIX Index as forward looking market data to compare the nonparametric filtered historical simulation, GARCH model with joint likelihood estimation (parametric), Restricted Boltzmann Machine and the conditional Variational Autoencoder (Generative Networks) for their suitability as PiT ESG. Our evaluation consists of statistical tests for model fit and benchmarking the out of sample forecasting quality with a strategy backtest using model output as stop loss criterion. We find that both Generative Networks outperform the nonparametric and classic parametric model in our tests, but that the CVAE seems to be particularly well suited for our purposes: yielding more robust performance and being computationally lighter.

Wang, Y., Polson, N. G., and Sokolov, V. O. (2019). “Scalable Data Augmentation for Deep Learning.” In: *arXiv e-Print*.

Scalable Data Augmentation (SDA) provides a framework for training deep learning models using auxiliary hidden layers. Scalable MCMC is available for network training and inference. SDA provides a number of computational advantages over traditional algorithms, such as avoiding backtracking, local modes and can perform optimization with stochastic gradient descent (SGD) in TensorFlow. Standard deep neural networks with logit, ReLU and SVM activation functions are straightforward to implement. To illustrate our architectures and methodology, we use Polya-Gamma logit data augmentation for a number of standard datasets. Finally, we conclude with directions for future research.

Wang, Y., Huang, G., Song, S., Pan, X., Xia, Y., and Wu, C. (2020c). “Regularizing Deep Networks with Semantic Data Augmentation.” In: *arXiv e-Print*.

Data augmentation is widely known as a simple yet surprisingly effective technique for regularizing deep networks. Conventional data augmentation schemes, e.g., flipping, translation or rotation, are low-level, data-independent and class-agnostic operations, leading to limited diversity for augmented samples. To this end, we propose a novel semantic data augmentation algorithm to complement traditional approaches. The proposed method is inspired by the intriguing property that deep networks are effective in learning linearized features, i.e., certain directions in the deep feature space correspond to meaningful semantic transformations, e.g., changing the background or view angle of an object. Based on this observation, translating training samples along many such directions in the feature space can effectively augment the dataset for more diversity. To implement this idea, we first introduce a sampling based method to obtain semantically meaningful directions efficiently. Then, an upper bound of the expected cross-entropy (CE) loss on the augmented training set is derived by assuming the number of augmented samples goes to infinity, yielding a highly efficient algorithm. In fact, we show that the proposed implicit semantic data augmentation (ISDA) algorithm amounts to minimizing a novel robust CE loss, which adds minimal extra computational cost to a normal training procedure. In addition to supervised learning, ISDA can be applied to semi-supervised learning tasks under the consistency regularization framework, where ISDA amounts to minimizing the upper bound of the expected KL-divergence between the augmented features and the original features. Although being simple, ISDA consistently improves the generalization performance of popular deep models (e.g., ResNets and DenseNets) on a variety of datasets, i.e., CIFAR-10, CIFAR-100, SVHN, ImageNet, Cityscapes and MS COCO.

Wen, Q., Sun, L., Yang, F., Song, X., Gao, J., Wang, X., and Xu, H. (2021). “Time Series Data Augmentation for Deep Learning: A Survey.” In: *Proceedings of the Thirtieth IJCAI Conference*.

Deep learning performs remarkably well on many time series analysis tasks recently. The superior performance of deep neural networks relies heavily on a large number of training data to avoid overfitting. However, the labeled data of many real-world time series applications may be limited such as classification in medical time series and anomaly detection in AIOps. As an effective way to enhance the size and quality of the training data, data augmentation is crucial to the successful application of deep learning models on time series data. In this paper, we systematically review different data augmentation methods for time series. We propose a taxonomy for the reviewed methods, and then provide a structured review for these methods by highlighting their strengths and limitations. We also empirically compare different data augmentation methods for different tasks including time series classification, anomaly detection, and forecasting. Finally, we discuss and highlight five future directions to provide useful research guidance.

Westphal, R. and Sornette, D. (2020). “How Market Intervention Can Prevent Bubbles and Crashes.” In: *SSRN e-Print*.

Using an agent-based model (ABM) with fundamentalists and chartists, prone to develop bubbles and crashes, we demonstrate the usefulness of direct market intervention by a policy maker, documenting strong performance in preventing bubbles and drawdowns and augmenting significantly the welfare of all investors. In our ABM, the policy maker diagnoses burgeoning bubbles by forming an expectation of the future return of the risky asset

in the form of an exponential moving average of the excess return over the long-term return. The policy maker invests in the risky asset when he detects a small deviation of the return from the long-term growth rate in order to construct an inventory that he draws upon later to fight future market exuberance. Then, when this deviation between the current growth rate and the long-term growth rate exceeds the policy maker's tolerance level, he starts to sell the risky asset that he has accumulated earlier, in a countercyclical fight against future price increase. We find that the policy maker succeeds in preventing bubbles and crashes in our ABM. In simulations without bubbles, the policy maker behaves similarly to the fundamentalists and his impact is negligible, following the principle of "Primum non nocere". In simulations where bubbles form spontaneously as a result of the noise traders's strategies, the policy maker's intervention reduces the average drawdown by a factor of two when his market impact becomes significant. We find that the policy maker intervention improves all analysed metrics of market returns, including volatility, skewness, kurtosis and VaR, making the market less turbulent and more stable. The combination of fewer bubbles and crashes, lower market risks and the stability of the long-term growth rate make the policy maker intervention to improve the welfare of all investors as measured by their risk-adjusted return, increasing the Sharpe ratios from approximately 0.3 to 0.5 for noise traders, from 0.6 to 0.8 for fundamentalists as the market impact of the policy maker increases to the level of the fundamentalists. We also test the sensitivity of these results to variations of the key parameters of the strategy of the policy maker and find very robust outcomes. In particular, the conclusions are unchanged even under very large miscalibrated long-term expected returns of the risky asset.

Wiese, M., Knobloch, R., Korn, R., and Kretschmer, P. (2020). "Quant GANs: deep generation of financial time series." In: *Quantitative Finance* 20(9), pp. 1419–1440.

Modeling financial time series by stochastic processes is a challenging task and a central area of research in financial mathematics. As an alternative, we introduce Quant GANs, a data-driven model which is inspired by the recent success of generative adversarial networks (GANs). Quant GANs consist of a generator and discriminator function, which utilize temporal convolutional networks (TCNs) and thereby achieve to capture long-range dependencies such as the presence of volatility clusters. The generator function is explicitly constructed such that the induced stochastic process allows a transition to its risk-neutral distribution. Our numerical results highlight that distributional properties for small and large lags are in an excellent agreement and dependence properties such as volatility clusters, leverage effects, and serial autocorrelations can be generated by the generator function of Quant GANs, demonstrably in high fidelity.

Wood, M. (2018). "How sure are we? Two approaches to statistical inference." In: *arXiv e-Print*.

Suppose you are told that taking a statin will reduce your risk of a heart attack or stroke by 3% in the next ten years, or that women have better emotional intelligence than men. You may wonder how accurate the 3% is, or how confident we should be about the assertion about women's emotional intelligence, bearing in mind that these conclusions are only based on samples of data? My aim here is to present two statistical approaches to questions like these. Approach 1 is often called null hypothesis testing but I prefer the phrase "baseline hypothesis": this is the standard approach in many areas of inquiry but is fraught with problems. Approach 2 can be viewed as a generalisation of the idea of confidence intervals, or as the application of Bayes' theorem. Unlike Approach 1, Approach 2 provides a tentative estimate of the probability of hypotheses of interest. For both approaches, I explain, from first principles, building only on "common sense" statistical concepts like averages and randomness, both how to derive answers, and the rationale behind the answers. This is achieved by using computer simulation methods (resampling and bootstrapping using a spreadsheet available on the web) which avoid the use of probability distributions (t, normal, etc). Such a minimalist, but reasonably rigorous, analysis is particularly useful in a discipline like statistics which is widely used by people who are not specialists. My intended audience includes both statisticians, and users of statistical methods who are not statistical experts.

Xie, Q., Dai, Z., Hovy, E., Luong, M.-T., and Le, Q. V. (2019). "Unsupervised Data Augmentation." In: *arXiv e-Print*.

Despite its success, deep learning still needs large labeled datasets to succeed. Data augmentation has shown much promise in alleviating the need for more labeled data, but it so far has mostly been applied in supervised settings and achieved limited gains. In this work, we propose to apply data augmentation to unlabeled data in a semi-supervised learning setting. Our method, named Unsupervised Data Augmentation or UDA, encourages the model predictions to be consistent between an unlabeled example and an augmented unlabeled example. Unlike previous methods that use random noise such as Gaussian noise or dropout noise, UDA has a small twist in that it makes use of harder and more realistic noise generated by state-of-the-art data augmentation methods. This small twist leads to substantial improvements on six language tasks and three vision tasks even when the labeled

set is extremely small. For example, on the IMDb text classification dataset, with only 20 labeled examples, UDA outperforms the state-of-the-art model trained on 25,000 labeled examples. On standard semi-supervised learning benchmarks, CIFAR-10 with 4,000 examples and SVHN with 1,000 examples, UDA outperforms all previous approaches and reduces more than 30% of the error rates of state-of-the-art methods: going from 7.66% to 5.27% and from 3.53% to 2.46% respectively. UDA also works well on datasets that have a lot of labeled data. For example, on ImageNet, with 1.3M extra unlabeled data, UDA improves the top-1/top-5 accuracy from 78.28/94.36% to 79.04/94.45% when compared to AutoAugment.

Xu, H., Wang, J., Li, H., Ouyang, D., and Shao, J. (2021). “Unsupervised meta-learning for few-shot learning.” In: *Pattern Recognition* 116, p. 107951.

Meta-learning is an effective tool to address the few-shot learning problem, which requires new data to be classified considering only a few training examples. However, when used for classification, it requires large labeled datasets, which are not always available in practice. In this paper, we propose an unsupervised meta-learning algorithm that learns from an unlabeled dataset and adapts to downstream human-specific tasks with few labeled data. The proposed algorithm constructs tasks using clustering embedding methods and data augmentation functions to satisfy two critical class distinction requirements. To alleviate the biases and the weak diversity problem introduced by data augmentation functions, the proposed algorithm uses two methods, which are shifting the feeding data between the inner-outer loops and a novel data augmentation function. We further provide theoretical analysis of the effect of augmentation data in the inner/outer loop. Experiments on the MiniImagenet and Omniglot datasets demonstrate that the proposed unsupervised meta-learning approach outperforms other tested unsupervised representation learning approaches and two recent unsupervised meta-learning baselines. Compared with supervised meta-learning approaches, certain results produced by our method are quite close to those produced by such methods trained on the human-designed labeled tasks.

Xu, L., Gotwalt, C., Hong, Y., King, C. B., and Meeker, W. Q. (2020). “Applications of the Fractional-Random-Weight Bootstrap.” In: *The American Statistician*, pp. 1–21.

For several decades, the resampling based bootstrap has been widely used for computing confidence intervals (CIs) for applications where no exact method is available. However, there are many applications where the resampling bootstrap method cannot be used. These include situations where the data are heavily censored due to the success response being a rare event, situations where there is insufficient mixing of successes and failures across the explanatory variable(s), and designed experiments where the number of parameters is close to the number of observations. These three situations all have in common that there may be a substantial proportion of the resamples where it is not possible to estimate all of the parameters in the model. This article reviews the fractional-random-weight bootstrap method and demonstrates how it can be used to avoid these problems and construct CIs in a way that is accessible to statistical practitioners. The fractional-random-weight bootstrap method is easy to use and has advantages over the resampling method in many challenging applications.

Yacoby, Y., Pan, W., and Doshi-Velez, F. (2020). “Characterizing and Avoiding Problematic Global Optima of Variational Autoencoders.” In: *Proceedings of Machine Learning Research*, pp. 1–17.

Variational Auto-encoders (VAEs) are deep generative latent variable models consisting of two components: a generative model that captures a data distribution  $p(x)$  by transforming a distribution  $p(z)$  over latent space, and an inference model that infers likely latent codes for each data point (Kingma and Welling, 2013). Recent work shows that traditional training methods tend to yield solutions that violate modeling desiderata: (1) the learned generative model captures the observed data distribution but does so while ignoring the latent codes, resulting in codes that do not represent the data (e.g. van den Oord et al. (2017); Kim et al. (2018)); (2) the aggregate of the learned latent codes does not match the prior  $p(z)$ . This mismatch means that the learned generative model will be unable to generate realistic data with samples from  $p(z)$  (e.g. Makhzani et al. (2015); Tomczak and Welling (2017)). In this paper, we demonstrate that both issues stem from the fact that the global optima of the VAE training objective often correspond to undesirable solutions. Our analysis builds on two observations: (1) the generative model is unidentifiable - there exist many generative models that explain the data equally well, each with different (and potentially unwanted) properties and (2) bias in the VAE objective - the VAE objective may prefer generative models that explain the data poorly but have posteriors that are easy to approximate. We present a novel inference method, LiBI, mitigating the problems identified in our analysis. On synthetic datasets, we show that LiBI can learn generative models that capture the data distribution and inference models that better satisfy modeling assumptions when traditional methods struggle to do so.

Yao, H., Jia, X., Kumar, V., and Li, Z. (2020). “Learning with Small Data.” In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM.

In the era of big data, data-driven methods have become increasingly popular in various applications, such as image recognition, traffic signal control, fake news detection. The superior performance of these data-driven approaches relies on large-scale labeled training data, which are probably inaccessible in real-world applications, i.e., "small (labeled) data" challenge. Examples include predicting emergent events in a city, detecting emerging fake news, and forecasting the progression of conditions for rare diseases. In most scenarios, people care about these small data cases most and thus improving the learning effectiveness of machine learning algorithms with small labeled data has been a popular research topic. In this tutorial, we will review the trending state-of-the-art machine learning techniques for learning with small (labeled) data. These techniques are organized from two aspects: (1) providing a comprehensive review of recent studies about knowledge generalization, transfer, and sharing, where transfer learning, multi-task learning, and meta-learning are discussed. Particularly, we will focus more on meta-learning, which improves the model generalization ability and has been proven to be an effective approach recently; (2) introducing the cutting-edge techniques which focus on incorporating domain knowledge into machine learning models. Different from model-based knowledge transfer techniques, in real-world applications, domain knowledge (e.g., physical laws) provides us with a new angle to deal with the small data challenge. Specifically, domain knowledge can be used to optimize learning strategies and/or guide the model design. In data mining field, we believe that learning with small data is a trending topic with important social impact, which will attract both researchers and practitioners from academia and industry.

Ye, R. and Dai, Q. (2021). "Implementing transfer learning across different datasets for time series forecasting." In: *Pattern Recognition* 109, p. 107617.

Due to the extensive practical value of time series prediction, many excellent algorithms have been proposed. Most of these methods are developed assuming that massive labeled training data are available. However, this assumption might be invalid in some actual situations. To address this limitation, a transfer learning framework with deep architectures is proposed. Since convolutional neural network (CNN) owns favorable feature extraction capability and can implement parallelization more easily, we propose a deep transfer learning method resorting to the architecture of CNN, termed as DTr-CNN for short. It can effectively alleviate the available labeled data absence and leverage useful knowledge to the current prediction. Notably, in our method, transfer learning process is implemented across different datasets. For a given target domain, in real-world scenarios, relativity of truly available potential source datasets may not be obvious, which is challenging and rarely referred to in most existing time series prediction methods. Aiming at this problem, the incorporation of Dynamic Time Warping (DTW) and Jensen-Shannon (JS) divergence is adopted for the selection of the appropriate source domain. Effectiveness of the proposed method is empirically underpinned by the experiments conducted on one group of synthetic and two groups of practical datasets. Besides, an additional experiment on NN5 dataset is conducted.

Yeh, C.-C. M. (2020). "Towards a Near Universal Time Series Data Mining Tool: Introducing the Matrix Profile." In: *arXiv e-Print*.

The last decade has seen a flurry of research on all-pairs-similarity-search (or, self-join) for text, DNA, and a handful of other datatypes, and these systems have been applied to many diverse data mining problems. Surprisingly, however, little progress has been made on addressing this problem for time series subsequences. In this thesis, we have introduced a near universal time series data mining tool called matrix profile which solves the all-pairs-similarity-search problem and caches the output in an easy-to-access fashion. The proposed algorithm is not only parameter-free, exact and scalable, but also applicable for both single and multidimensional time series. By building time series data mining methods on top of matrix profile, many time series data mining tasks (e.g., motif discovery, discord discovery, shapelet discovery, semantic segmentation, and clustering) can be efficiently solved. Because the same matrix profile can be shared by a diverse set of time series data mining methods, matrix profile is versatile and computed-once-use-many-times data structure. We demonstrate the utility of matrix profile for many time series data mining problems, including motif discovery, discord discovery, weakly labeled time series classification, and representation learning on domains as diverse as seismology, entomology, music processing, bioinformatics, human activity monitoring, electrical power-demand monitoring, and medicine. We hope the matrix profile is not the end but the beginning of many more time series data mining projects.

Yeh, C.-C. M., Zhu, Y., Ulanova, L., Begum, N., Ding, Y., Dau, H. A., Zimmerman, Z., Silva, D. F., Mueen, A., and Keogh, E. (2017). "Time series joins, motifs, discords and shapelets: a unifying view that exploits the matrix profile." In: *Data Mining and Knowledge Discovery* 32(1), pp. 1–41.

The last decade has seen a flurry of research on all-pairs-similarity-search (or similarity joins) for text, DNA and a handful of other datatypes, and these systems have been applied to many diverse data mining problems. However, there has been surprisingly little progress made on similarity joins for time series subsequences. The lack

of progress probably stems from the daunting nature of the problem. For even modest sized datasets the obvious nested-loop algorithm can take months, and the typical speed-up techniques in this domain (i.e., indexing, lower-bounding, triangular-inequality pruning and early abandoning) at best produce only one or two orders of magnitude speedup. In this work we introduce a novel scalable algorithm for time series subsequence all-pairs-similarity-search. For exceptionally large datasets, the algorithm can be trivially cast as an anytime algorithm and produce high-quality approximate solutions in reasonable time and/or be accelerated by a trivial porting to a GPU framework. The exact similarity join algorithm computes the answer to the time series motif and time series discord problem as a side-effect, and our algorithm incidentally provides the fastest known algorithm for both these extensively-studied problems. We demonstrate the utility of our ideas for many time series data mining problems, including motif discovery, novelty discovery, shapelet discovery, semantic segmentation, density estimation, and contrast set mining. Moreover, we demonstrate the utility of our ideas on domains as diverse as seismology, music processing, bioinformatics, human activity monitoring, electrical power-demand monitoring and medicine.

Yilmaz, A., Kara, M., and Ozdemir, O. (2021). “Comparison of different estimation methods for extreme value distribution.” In: *Journal of Applied Statistics* 48(13-15), pp. 2259–2284.

The extreme value distribution was developed for modeling extreme-order statistics or extreme events. In this study, we discuss the distribution of the largest extreme. The main objective of this paper is to determine the best estimators of the unknown parameters of the extreme value distribution. Thus, both classical and Bayesian methods are used. The classical estimation methods under consideration are maximum likelihood estimators, moment’s estimators, least squares estimators, and weighted least squares estimators, percentile estimators, the ordinary least squares estimators, best linear unbiased estimators, L-moments estimators, trimmed L-moments estimators, and Bain and Engelhardt estimators. We also propose new estimators for the unknown parameters. Bayesian estimators of the parameters are derived by using Lindley’s approximation and Markov Chain Monte Carlo methods. The asymptotic confidence intervals are considered by using maximum likelihood estimators. The Bayesian credible intervals are also obtained by using Gibbs sampling. The performances of these estimation methods are compared with respect to their biases and mean square errors through a simulation study. The maximum daily flood discharge (annual) data sets of the Meric River and Feather River are analyzed at the end of the study for a better understanding of the methods presented in this paper.

Yoon, J., Jarrett, D., and van der Schaar, M. (2019). “Time-series Generative Adversarial Networks.” In: *NIPS*. NIPS.

A good generative model for time-series data should preserve temporal dynamics, in the sense that new sequences respect the original relationships between variables across time. Existing methods that bring generative adversarial networks (GANs) into the sequential setting do not adequately attend to the temporal correlations unique to time-series data. At the same time, supervised models for sequence prediction - which allow finer control over network dynamics - are inherently deterministic. We propose a novel framework for generating realistic time-series data that combines the flexibility of the unsupervised paradigm with the control afforded by supervised training. Through a learned embedding space jointly optimized with both supervised and adversarial objectives, we encourage the network to adhere to the dynamics of the training data during sampling. Empirically, we evaluate the ability of our method to generate realistic samples using a variety of real and synthetic time-series datasets. Qualitatively and quantitatively, we find that the proposed framework consistently and significantly outperforms state-of-the-art benchmarks with respect to measures of similarity and predictive ability.

Yu, P. L. H., Ng, F. C., and Ting, J. K. W. (2020). “Adjusting covariance matrix for risk management.” In: *Quantitative Finance* 20(10), pp. 1681–1699.

The covariance matrix of asset returns can change drastically and generate huge losses in portfolio value under extreme conditions such as market interventions and financial crises. Estimation of the covariance matrix under a chaotic market is often a call to action in risk management. Nowadays, stress testing has become a standard procedure for many financial institutions to estimate the capital requirement for their portfolio holdings under various stress scenarios. A possible stress scenario is to adjust the covariance matrix to mimic the situation under an underlying stress event. It is reasonable that when some covariances are altered, other covariances should vary as well. Recently, Ng et al. proposed a unified approach to determine a proper correlation matrix which reflects the subjective views of correlations. However, this approach requires matrix vectorization and hence it is not computationally efficient for high dimensional matrices. Besides, it only adjusts correlations, but it is well known that high correlations often go together with high standard deviations during a crisis period. To address



these limitations, we propose a Bayesian approach to covariance matrix adjustment by incorporating subjective views of covariances. Our approach is computationally efficient and can be applied to high dimensional matrices.

Yu, R. (2020). “A Tutorial on VAEs: From Bayes’ Rule to Lossless Compression.” In: *arXiv e-Print*.

The Variational Auto-Encoder (VAE) is a simple, efficient, and popular deep maximum likelihood model. Though usage of VAEs is widespread, the derivation of the VAE is not as widely understood. In this tutorial, we will provide an overview of the VAE and a tour through various derivations and interpretations of the VAE objective. From a probabilistic standpoint, we will examine the VAE through the lens of Bayes’ Rule, importance sampling, and the change-of-variables formula. From an information theoretic standpoint, we will examine the VAE through the lens of lossless compression and transmission through a noisy channel. We will then identify two common misconceptions over the VAE formulation and their practical consequences. Finally, we will visualize the capabilities and limitations of VAEs using a code example (with an accompanying Jupyter notebook) on toy 2D data.

Yuan, J. and Yuan, X. (2021). “A Monte Carlo synthetic sample based performance evaluation method for covariance matrix estimators.” In: *Applied Economics Letters*.

The evaluation of covariance matrix estimators is very important for portfolio analysis and risk management. The Monte Carlo synthetic sample based performance evaluation method proposed by this article can avoid the main shortcomings of statistical and economic methods which are widely used in the existing literature. The proposed method does not need the true covariance and does not need to introduce the performance of the out-of-sample portfolios. It is an intuitive, effective and robust measure for both simulation and empirical analysis.

Zhang, J. (2020). “Modern Monte Carlo methods for efficient uncertainty quantification and propagation: A survey.” In: *WIREs Computational Statistics* 13(5).

Uncertainty quantification (UQ) includes the characterization, integration, and propagation of uncertainties that result from stochastic variations and a lack of knowledge or data in the natural world. Monte Carlo (MC) method is a sampling-based approach that has widely used for quantification and propagation of uncertainties. However, the standard MC method is often time-consuming if the simulation-based model is computationally intensive. This article gives an overview of modern MC methods to address the existing challenges of the standard MC in the context of UQ. Specifically, multilevel Monte Carlo (MLMC) extending the concept of control variates achieves a significant reduction of the computational cost by performing most evaluations with low accuracy and corresponding low cost, and relatively few evaluations at high accuracy and corresponding high cost. Multifidelity Monte Carlo (MFMC) accelerates the convergence of standard Monte Carlo by generalizing the control variates with different models having varying fidelities and varying computational costs. Multimodel Monte Carlo method (MMMC), having a different setting of MLMC and MFMC, aims to address the issue of UQ and propagation when data for characterizing probability distributions are limited. Multimodel inference combined with importance sampling is proposed for quantifying and efficiently propagating the uncertainties resulting from small data sets. All of these three modern MC methods achieve a significant improvement of computational efficiency for probabilistic UQ, particularly uncertainty propagation. An algorithm summary and the corresponding code implementation are provided for each of the modern MC methods. The extension and application of these methods are discussed in detail.

Zhang, S. (2018). “Optimal Retirement Planning: Scenario Generation, Preferences, and Objectives.” PhD thesis. University of Waterloo.

The global trend of shifting from defined benefit (DB) to defined contribution (DC) workplace pension plans is putting growing pressure on individuals to take more ownership in retirement planning and financial decision-making. The essence of the DB is the life-long income guarantee, which requires limited financial planning decisions to be made, either in the accumulation or decumulation phase. The DC on the other hand, is significantly more complex. The lump sum payment at retirement burdens individuals with the task of income generation, in the presence of challenges stemming from an uncertain future lifetime, economic conditions, and evolving consumption needs. The average retiree has limited competency to navigate these challenges, due to low financial literacy, lack of willpower, or deteriorating cognitive abilities with older ages. The high stake of these challenges calls for a normative solution to be proposed - a solution that considers the intricacy of risks, preferences, and normative objective formulations. The objective of this thesis is to explore such a solution. This thesis comprises three inter-related research directions: long-term economic scenario generators (ESGs), recursive preferences in life-cycle portfolio selection, and retirement objective formulation. A brief description of the subsequent chapters will now follow. The first chapter conducts a review of Wilkie’s ESG, with analysis

restricted to series pertinent to retirement planning. Our main findings indicate that there exist challenges in modelling long-term economic series due to the presence of multiple structural shifts in the historical time series. Consequently, certain assumptions of stationarity are violated, and parameters are sensitive to the calibration period. A backtest based on 30-year out-of-sample data indicated that over that period the model had tended to overestimate inflation, underestimate total return on stocks, and performed relatively well for long-term interest rates. Additionally, Wilkie’s ESG can be under-representative of the risk in long-term stock investment, particularly in the tails. The second chapter provides an introductory discussion of Epstein-Zin preferences, which are adopted in the succeeding chapter as a normative preference model. The purpose is to first investigate the implied optimal behaviour and its plausibility. We pay particular attention to whether the output leads to plausible behaviour given the context of retirement planning. Specifically, analytical solutions for a simple consumption problem are derived, isolating the impact of relative risk aversion (RRA), elasticity of intertemporal substitution (EIS), time discounting, and risks stemming from mortality, investment, and inflation. We investigate three Epstein-Zin models employed in the literature, which differ in their treatment of mortality risk, and find that some lead to normatively implausible solutions. Importantly, we find that the EIS is not always monotone in its effect on consumption volatility over time, meaning that its interpretation can be ambiguous when considering an uncertain future lifetime. This has been misinterpreted in the literature to date. We also show that one particular Epstein-Zin specification is not necessarily a generalization of expected utility maximization under constant relative risk aversion, as many works wrongly claim. The third chapter investigates the normative validity of the optimal consumption and investment strategies of a discrete-time Epstein-Zin utility maximizing DC retiree who wishes to benefit from stock investment, longevity insurance, and inflation protection. A comparison of three Epstein-Zin specifications is conducted. We use a combination of qualitative and quantitative criteria to evaluate the adequacy of the optimal consumption profile, with special attention paid to the downside risk at extreme old ages. We find that it remains optimal to fully annuitize, but agents with high relative risk aversion hold precautionary savings, the level of which is impacted by the EIS and the preference specification. As discussed in the preceding chapter, the interpretation of EIS on consumption volatility is found to be ambiguous. Investigations of the optimal consumption profile reveal that agents are exposed to relatively high levels of downside risk in the long run. This is partially attributed to a time discounting factor less than 1, which implicitly (and contradictorily) assumes myopia in normative decision-making. An investigation of zero time discounting is conducted, with downside risk found to be significantly reduced in the long run. The fourth chapter focuses on retirement objective formulation. This chapter is motivated by the unsatisfactory normative solutions found in the preceding chapter under mathematically convenient objective functions. In order to develop more actionable prescriptive solutions, we seek to holistically explore actual retirement decision-making. To this end, we conduct a survey study of 1,000 Canadian (pre-)retirees age 50 to 80, on topics of retirement consumption, wealth, income, risk perception, decision making, and planning objectives. Additionally, we investigate the descriptive validity of the expected lifetime discounted utility maximization framework in predicting optimal planning behaviours. Overall, there is overwhelming evidence of heterogeneity in wealth, income, concerns, and objectives. We find a prevalence of low retirement assets, a severe underestimation of survival probabilities to an extreme old age of 95, and a strong aversion toward life annuities. Pre-retirees appear to have reasonable expectations regarding income and assets in retirement, with the median retiree relying heavily on public pension sources. (Pre-)retirees are primarily concerned with liquidity needs, consumption smoothing, inflation, and longevity in retirement, and are least concerned with bequests. We elicited risk and time preferences, and found an average RRA parameter between 1.74 to 2.48 for pre-retirees and 2.48 to 3.74 for retirees, and a median subjective time discount factor of 0.997. A study of decision-making under risky scenarios reveals dramatic differences between the actual and implied choices under the expected utility maximization framework. Particularly, in the presence of inflation risk, agents lack the understanding of the long-term cumulative impact of inflation on the cost of living. In the presence of investment risk, the upside gain drives decision-making, and the presence of minimum income protection effectively provided by public pension income induces more risk-taking behaviour. The last chapter concludes the thesis, and proposes general directions for future work in retirement planning research.

Zhang, X., Wang, Z., Liu, D., and Ling, Q. (2018). “DADA: Deep Adversarial Data Augmentation for Extremely Low Data Regime Classification.” In: *arXiv e-Print*.

Deep learning has revolutionized the performance of classification, but meanwhile demands sufficient labeled data for training. Given insufficient data, while many techniques have been developed to help combat overfitting, the challenge remains if one tries to train deep networks, especially in the ill-posed extremely low data regimes: only a small set of labeled data are available, and nothing – including unlabeled data – else. Such regimes arise from

practical situations where not only data labeling but also data collection itself is expensive. We propose a deep adversarial data augmentation (DADA) technique to address the problem, in which we elaborately formulate data augmentation as a problem of training a class-conditional and supervised generative adversarial network (GAN). Specifically, a new discriminator loss is proposed to fit the goal of data augmentation, through which both real and augmented samples are enforced to contribute to and be consistent in finding the decision boundaries. Tailored training techniques are developed accordingly. To quantitatively validate its effectiveness, we first perform extensive simulations to show that DADA substantially outperforms both traditional data augmentation and a few GAN-based options. We then extend experiments to three real-world small labeled datasets where existing data augmentation and/or transfer learning strategies are either less effective or infeasible. All results endorse the superior capability of DADA in enhancing the generalization ability of deep networks trained in practical extremely low data regimes. Source code is available at <https://github.com/SchafferZhang/DADA>.

Zhao, L. (2022). “Event Prediction in the Big Data Era.” In: *ACM Computing Surveys* 54(5), pp. 1–37.

Events are occurrences in specific locations, time, and semantics that nontrivially impact either our society or the nature, such as earthquakes, civil unrest, system failures, pandemics, and crimes. It is highly desirable to be able to anticipate the occurrence of such events in advance to reduce the potential social upheaval and damage caused. Event prediction, which has traditionally been prohibitively challenging, is now becoming a viable option in the big data era and is thus experiencing rapid growth, also thanks to advances in high performance computers and new Artificial Intelligence techniques. There is a large amount of existing work that focuses on addressing the challenges involved, including heterogeneous multi-faceted outputs, complex (e.g., spatial, temporal, and semantic) dependencies, and streaming data feeds. Due to the strong interdisciplinary nature of event prediction problems, most existing event prediction methods were initially designed to deal with specific application domains, though the techniques and evaluation procedures utilized are usually generalizable across different domains. However, it is imperative yet difficult to cross-reference the techniques across different domains, given the absence of a comprehensive literature survey for event prediction. This article aims to provide a systematic and comprehensive survey of the technologies, applications, and evaluations of event prediction in the big data era. First, systematic categorization and summary of existing techniques are presented, which facilitate domain experts’ searches for suitable techniques and help model developers consolidate their research at the frontiers. Then, comprehensive categorization and summary of major application domains are provided to introduce wider applications to model developers to help them expand the impacts of their research. Evaluation metrics and procedures are summarized and standardized to unify the understanding of model performance among stakeholders, model developers, and domain experts in various application domains. Finally, open problems and future directions are discussed. Additional resources related to event prediction are included in the paper website.

Zhu, Y., Mariani, G., and Li, J. (2021). “Pagan: Portfolio Analysis with Generative Adversarial Networks.” In: *SSRN e-Print*.

Sound financial theories demonstrate that in an efficient marketplace all information available today, including expectations on future events, are represented in today prices whereas future price trend is driven by the uncertainty. This jeopardizes the efforts put in designing prediction models. To deal with the unpredictability of financial systems, today’s portfolio management is largely based on the Markowitz framework which puts more emphasis in the analysis of the market uncertainty and less in the price prediction. The limitation of the Markowitz framework stands in taking very strong ideal assumptions about future returns probability distribution. To address this situation we propose PAGAN, a pioneering methodology for portfolio management based on deep generative models. The goal is modeling the market uncertainty that ultimately is the main factor driving future trends. The generative model learns the joint probability distribution of price trends for a set of financial assets to match the probability distribution of the real market. Once the model is trained, a portfolio is optimized by deciding the best diversification to minimize the risk and maximize the expected returns observed over the execution of several simulations. Applying the model for analyzing possible futures, is as simple as executing a Monte Carlo simulation, a technique very familiar to finance experts. The experimental results on different portfolios representing different geopolitical areas and industrial segments constructed using real-world public data sets demonstrate promising results.

Zhu, Y., Imamura, M., Nikovski, D., and Keogh, E. (2018). “Time series chains: A novel tool for time series data mining.” In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. Ed. by J. Lang. California: International Joint Conferences on Artificial Intelligence Organization, pp. 5414–5418.

Since their introduction over a decade ago, time series motifs have become a fundamental tool for time series analytics, finding diverse uses in dozens of domains. In this work we introduce Time Series Chains, which are related to, but distinct from, time series motifs. Informally, time series chains are a temporally ordered set of subsequence patterns, such that each pattern is similar to the pattern that preceded it, but the first and last patterns are arbitrarily dissimilar. In the discrete space, this is similar to extracting the text chain "hit, hot, dot, dog" from a paragraph. The first and last words have nothing in common, yet they are connected by a chain of words with a small mutual difference. Time Series Chains can capture the evolution of systems, and help predict the future. As such, they potentially have implications for prognostics. In this work, we introduce a robust definition of time series chains, and a scalable algorithm that allows us to discover them in massive datasets.

Zhu, Y., Imamura, M., Nikovski, D., and Keogh, E. (2019). "Introducing time series chains: a new primitive for time series data mining." In: *Knowledge and Information Systems* 60, pp. 1135–1161.

Time series motifs were introduced in 2002 and have since become a fundamental tool for time series analytics, finding diverse uses in dozens of domains. In this work, we introduce Time Series Chains, which are related to, but distinct from, time series motifs. Informally, time series chains are a temporally ordered set of subsequence patterns, such that each pattern is similar to the pattern that preceded it, but the first and last patterns can be arbitrarily dissimilar. In the discrete space, this is similar to extracting the text chain , date, cate, cade, code from text stream. The first and last words have nothing in common, yet they are connected by a chain of words with a small mutual difference. Time series chains can capture the evolution of systems, and help predict the future. As such, they potentially have implications for prognostics. In this work, we introduce two robust definitions of time series chains and scalable algorithms that allow us to discover them in massive complex datasets.

Ziyin, L., Minami, K., and Imajo, K. (2021). "What Data Augmentation Do We Need for Deep-Learning-Based Finance?" In: *arXiv e-Print*.

The main task we consider is portfolio construction in a speculative market, a fundamental problem in modern finance. While various empirical works now exist to explore deep learning in finance, the theory side is almost non-existent. In this work, we focus on developing a theoretical framework for understanding the use of data augmentation for deep-learning-based approaches to quantitative finance. The proposed theory clarifies the role and necessity of data augmentation for finance; moreover, our theory motivates a simple algorithm of injecting a random noise of strength  $\sqrt{|r_{t-1}|}$  to the observed return  $r_t$ . This algorithm is shown to work well in practice.

Zorn, J. (2019). "Panic-aware portfolio optimization." In: *Journal of Asset Management*.

This article provides a portfolio optimization approach that takes into account extreme events. By merging a (downside-only) panic copula with the empirical marginal distributions, panic-awareness is attained for the optimization process. This approach includes the likelihood of highly co-dependent asset movements in panic states of the market empirically observed during market crashes. Panic-awareness CVaR optimization translates into robust equity portfolios, empirically exemplified for the UK and German stock market.